

Traitement du signal et Apprentissage profond ; applications industrielles

Thomas COURTAT

thomas.courtat@thalesgroup.com

Master MVA 2023

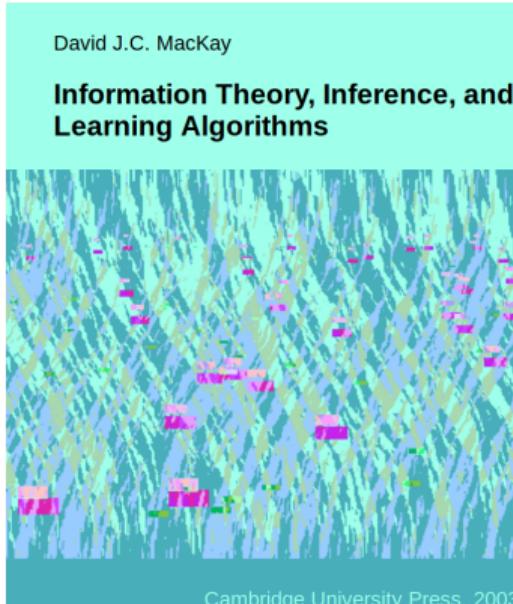
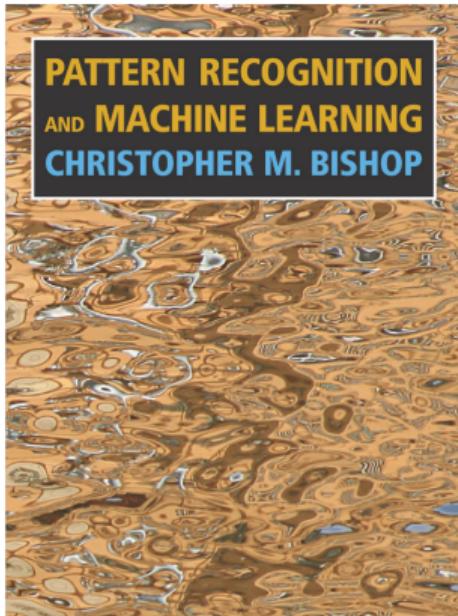


1 Estimation et Inférence

- Maximum de vraisemblance
- Modèles de mélange et classification
- Modèles Markoviens
- Inférence Bayésienne

2 Apprentissage Machine

Références



Cambridge University Press, 2003

Estimation et Inférence

Modèle statistique - Estimateurs

On modélise un phénomène (complexe) par une variable aléatoire X d'un espace Ω vers $\mathcal{E} (= \mathbb{R}, \mathbb{R}^K \dots)$ et de densité de probabilité $p_{\theta_*} \in \mathcal{P} = \{p_{\theta}, \theta \in \Theta\}$.

Modèle statistique - Estimateurs

On modélise un phénomène (complexe) par une variable aléatoire X d'un espace Ω vers $\mathcal{E} (= \mathbb{R}, \mathbb{R}^K \dots)$ et de densité de probabilité $p_{\theta_*} \in \mathcal{P} = \{p_\theta, \theta \in \Theta\}$.

L'estimation statistique consiste à produire un estimateur de θ_* à partir de N réalisations de variables aléatoires X_i distribuées comme X :

$$\hat{\theta}_* \simeq \hat{\theta}_*(X_0, X_1, \dots, X_{N-1})$$

Modèle statistique - Estimateurs

On modélise un phénomène (complexe) par une variable aléatoire X d'un espace Ω vers $\mathcal{E} (= \mathbb{R}, \mathbb{R}^K \dots)$ et de densité de probabilité $p_{\theta_*} \in \mathcal{P} = \{p_{\theta}, \theta \in \Theta\}$.

L'estimation statistique consiste à produire un estimateur de θ_* à partir de N réalisations de variables aléatoires X_i distribuées comme X :

$$\hat{\theta}_* \simeq \hat{\theta}_*(X_0, X_1, \dots, X_{N-1})$$

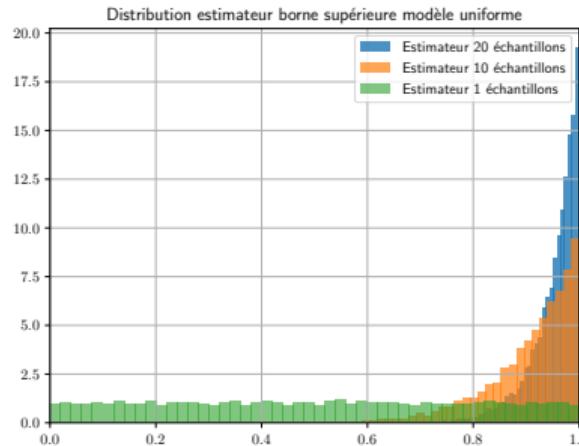
Exemples:

- $p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, $\theta = (\mu, \sigma^2)$, $\hat{\mu} = \frac{1}{n} \sum X_i$, $\hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \hat{\mu})^2$
- $p_{\theta}(x) = \frac{1}{M} \mathbf{1}_{[0,M]}(x)$, $\theta = (M)$, $\hat{M} = \max X_i$

Modèle statistique - Estimateurs

Un estimateur est une variable aléatoire avec une distribution, un espérance...

$$p_{\theta}(x) = \frac{1}{M} \mathbf{1}_{[0,M]}(x), \theta = (M), \quad \hat{M} = \max X_i$$



Maximum de vraisemblance

Maximum de vraisemblance

⇒ Méthode assez systématique pour trouver un estimateur de paramètres statistiques:

Maximum de vraisemblance

⇒ Méthode assez systématique pour trouver un estimateur de paramètres statistiques:

Sous des conditions assez générales, si on a une suite de variables aléatoires indépendantes, identiquement distribuées $X_i \sim p_{\theta_*}$, alors

$$\hat{\theta}_n = \operatorname{argmax} \frac{1}{n} \sum \log p_{\theta}(X_i)$$

est un estimateur asymptotiquement sans biais de θ_*

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV de μ pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$ avec σ^2 supposé connu.

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV de μ pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$ avec σ^2 supposé connu.

$$\frac{1}{n} \sum \log p_\mu(x_i) = \frac{1}{n} \sum (-\log \sqrt{2\pi\sigma^2}) - \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial \frac{1}{n} \sum \log p_\mu(x_i)}{\partial \mu} = -\frac{1}{n} \sum \frac{(x_i - \mu)}{\sigma^2}$$

$$\frac{\partial \frac{1}{n} \sum \log p_\mu(x_i)}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \frac{\sum x_i}{n}$$

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV de σ^2 pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$ avec μ supposé connu.

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV de σ^2 pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$ avec μ supposé connu.

On pose $s = \sigma^2$

$$\frac{1}{n} \sum \log p_s(x_i) = \frac{1}{n} \left(-\log \sqrt{2\pi s} - \frac{(x_i - \mu)^2}{2s} \right)$$

$$\frac{\partial \frac{1}{n} \sum \log p_s(x_i)}{\partial s} = \frac{1}{n} \sum \left(-\frac{1}{2s} + \frac{(x_i - \mu)^2}{2s^2} \right)$$

$$\frac{\partial \frac{1}{n} \sum \log p_s(x_i)}{\partial s} = 0 \Rightarrow \hat{s} = \frac{1}{n} \sum (x_i - \mu)^2$$

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV conjoint de (μ, σ^2) pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$.

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV conjoint de (μ, σ^2) pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$.

$$\frac{\partial \frac{1}{n} \sum \log p_{\mu, s}(x_i)}{\partial \mu} = -\frac{1}{n} \sum \frac{(x_i - \mu)}{s}$$

$$\frac{\partial \frac{1}{n} \sum \log p_{\mu, s}(x_i)}{\partial s} = \frac{1}{n} \sum \left(-\frac{1}{2s} + \frac{(x_i - \mu)^2}{2s^2} \right)$$

Annulation du gradient $\Rightarrow \hat{\mu} = \frac{\sum x_i}{n}$ et $\hat{s} = \frac{1}{n} \sum (x_i - \hat{\mu})^2$

Maximum de vraisemblance

⇒ Méthode assez systématique pour trouver un estimateur de paramètres statistiques:

Maximum de vraisemblance

⇒ Méthode assez systématique pour trouver un estimateur de paramètres statistiques:

Sous des conditions assez générales, si on a une suite de variables aléatoires indépendantes, identiquement distribuées $X_i \sim p_{\theta_*}$, alors

$$\hat{\theta}_n = \operatorname{argmax} \frac{1}{n} \sum \log p_{\theta}(X_i)$$

est un estimateur asymptotiquement sans biais de θ_*

Maximum de vraisemblance

⇒ Méthode assez systématique pour trouver un estimateur de paramètres statistiques:

Sous des conditions assez générales, si on a une suite de variables aléatoires indépendantes, identiquement distribuées $X_i \sim p_{\theta_*}$, alors

$$\hat{\theta}_n = \operatorname{argmax} \frac{1}{n} \sum \log p_{\theta}(X_i)$$

est un estimateur asymptotiquement sans biais de θ_*

de plus,

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow \mathcal{N}(0, F_{\{p_{\theta}, \theta \in \Theta\}}^{-1}(\theta_*))$$

avec $F_{\{p_{\theta}, \theta \in \Theta\}} = \mathbb{E}\left(-\frac{\partial^2 \log p_{\theta}(X_0)}{\partial \theta^2}\right)$ (*Information de Fisher*)

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV de μ et son information de Fisher pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$ avec σ^2 supposé connu.

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV de μ et son information de Fisher pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$ avec σ^2 supposé connu.

$$\frac{1}{n} \sum \log p_\mu(x_i) = \frac{1}{n} \sum (-\log \sqrt{2\pi\sigma^2}) - \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial \frac{1}{n} \sum \log p_\mu(x_i)}{\partial \mu} = -\frac{1}{n} \sum \frac{(x_i - \mu)}{\sigma^2}$$

$$\frac{\partial \frac{1}{n} \sum \log p_\mu(x_i)}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \frac{\sum x_i}{n}$$

$$\mathbb{E}\left(-\frac{\partial^2 \log p_\mu(x)}{\partial \mu^2}\right) = \frac{1}{\sigma^2}$$

$$\Rightarrow \sqrt{n}(\hat{\mu} - \mu) \rightarrow \mathcal{N}(0, \sigma^2)$$

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV de σ^2 et sa variance pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$ avec μ supposé connu.

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV de σ^2 et sa variance pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$ avec μ supposé connu.

On pose $s = \sigma^2$

$$\frac{1}{n} \sum \log p_s(x_i) = \frac{1}{n} \left(-\log \sqrt{2\pi s} - \frac{(x_i - \mu)^2}{2s} \right)$$

$$\frac{\partial \frac{1}{n} \sum \log p_s(x_i)}{\partial s} = \frac{1}{n} \sum \left(-\frac{1}{2s} + \frac{(x_i - \mu)^2}{2s^2} \right)$$

$$\frac{\partial \frac{1}{n} \sum \log p_s(x_i)}{\partial s} = 0 \Rightarrow \hat{s} = \frac{1}{n} \sum (x_i - \mu)^2$$

$$\mathbb{E}\left(-\frac{\partial^2 \log p_s(x)}{\partial s^2}\right) = -\frac{1}{2s^2} + \frac{2\mathbb{E}((x - \mu)^2)}{2s^3}$$

$$\mathbb{E}\left(-\frac{\partial^2 \log p_s(x)}{\partial s^2}\right) = -\frac{1}{2s^2} + \frac{1}{s^2}$$

$$\Rightarrow \sqrt{n}(\hat{s} - s) \rightarrow \mathcal{N}(0, 2s^4)$$

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV conjoint de (μ, σ^2) et sa variance pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$.

Maximum de vraisemblance - exemples

? Trouver l'estimateur MV conjoint de (μ, σ^2) et sa variance pour n variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$.

$$\frac{\partial \frac{1}{n} \sum \log p_{\mu,s}(x_i)}{\partial \mu} = -\frac{1}{n} \sum \frac{(x_i - \mu)}{s}$$

$$\frac{\partial \frac{1}{n} \sum \log p_{\mu,s}(x_i)}{\partial s} = \frac{1}{n} \sum \left(-\frac{1}{2s} + \frac{(x_i - \mu)^2}{2s^2} \right)$$

Annulation du gradient $\Rightarrow \hat{\mu} = \frac{1}{n} \sum x_i$ et $\hat{s} = \frac{1}{n} \sum (x_i - \hat{\mu})^2$

$$\mathbb{E}\left(-\frac{\partial^2 \log p_{\mu,s}(x)}{\partial \mu^2}\right) = \frac{1}{s}$$

$$\mathbb{E}\left(-\frac{\partial^2 \log p_{\mu,s}(x)}{\partial s^2}\right) = -\frac{1}{2s^2} + \frac{1}{s^2}$$

$$\mathbb{E}\left(-\frac{\partial^2 \log p_{\mu,s}(x)}{\partial \mu \partial s}\right) = \mathbb{E}\left(-\frac{x - \mu}{s^2}\right) = 0$$

$$\Rightarrow \sqrt{n} \left(\begin{pmatrix} \hat{\mu} \\ \hat{s}^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \right) \rightarrow \mathcal{N} \left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix} \right)$$

Maximum de vraisemblance - idée de preuve

Sous des conditions assez générales, si on a une suite de variables aléatoires indépendantes, identiquement distribuées $X_i \sim p_{\theta_*}$, alors

$$\hat{\theta}_n = \operatorname{argmax} \frac{1}{n} \sum \log p_{\theta}(X_i)$$

est un estimateur asymptotiquement sans biais de θ_*

de plus,

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow \mathcal{N}(0, F_{\{p_{\theta}, \theta \in \Theta\}}^{-1}(\theta_*))$$

avec $F_{\{p_{\theta}, \theta \in \Theta\}} = \mathbb{E}\left(-\frac{\partial^2 \log p_{\theta}(X_0)}{\partial \theta^2}\right)$ (*Information de Fisher*)

Maximum de vraisemblance - idée de preuve

- Si $X \sim p_{\theta_*}$ alors, $\theta_* \in \text{Argmax } \mathbb{E}(\log p_{\theta}(X))$

Maximum de vraisemblance - idée de preuve

- Si $X \sim p_{\theta_*}$ alors, $\theta_* \in \text{Argmax } \mathbb{E}(\log p_\theta(X))$

$$\begin{aligned}\mathbb{E}(\log \frac{p_\theta(X)}{p_{\theta_*}(X)}) &\leq \mathbb{E}(\frac{p_\theta(X)}{p_{\theta_*}(X)}) - 1 \\ &= \int \frac{p_\theta(X)}{p_{\theta_*}(X)} p_{\theta_*}(X) dX - 1 = 0\end{aligned}$$

Maximum de vraisemblance - idée de preuve

- Si $X \sim p_{\theta_*}$ alors, $\theta_* \in \text{Argmax } \mathbb{E}(\log p_{\theta}(X))$
- Si (X_i) iid comme X , $\frac{1}{n} \sum \log p_{\theta}(X_i) \rightarrow \mathbb{E}(\log p_{\theta}(X))$, $\forall \theta$

Maximum de vraisemblance - idée de preuve

- Si $X \sim p_{\theta_*}$ alors, $\theta_* \in \text{Argmax } \mathbb{E}(\log p_{\theta}(X))$
- Si (X_i) iid comme X , $\frac{1}{n} \sum \log p_{\theta}(X_i) \rightarrow \mathbb{E}(\log p_{\theta}(X))$, $\forall \theta$
- Si $\hat{\theta}_n = \text{argmax } \frac{1}{n} \sum \log p_{\theta}(X_i)$ alors, $\hat{\theta}_n \rightarrow \theta_*$

Maximum de vraisemblance - idée de preuve

- Si $X \sim p_{\theta_*}$ alors, $\theta_* \in \text{Argmax } \mathbb{E}(\log p_{\theta}(X))$
- Si (X_i) iid comme X , $\frac{1}{n} \sum \log p_{\theta}(X_i) \rightarrow \mathbb{E}(\log p_{\theta}(X))$, $\forall \theta$
- Si $\hat{\theta}_n = \text{argmax } \frac{1}{n} \sum \log p_{\theta}(X_i)$ alors, $\hat{\theta}_n \rightarrow \theta_*$
- $\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow \mathcal{N}(0, F^{-1}(\theta_*))$

Maximum de vraisemblance - idée de preuve

- $\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow \mathcal{N}(0, F^{-1}(\theta_*))$

Trois expressions équivalentes pour l'information de Fisher:

- $F_{\{p_\theta, \theta \in \Theta\}} = \mathbb{E}\left(-\frac{\partial^2 \log p_\theta(X_0)}{\partial \theta^2}\right)$
- $F_{\{p_\theta, \theta \in \Theta\}} = \mathbb{E}\left(\left(\frac{\partial \log p_\theta(X_0)}{\partial \theta}\right)^2\right)$
- $F_{\{p_\theta, \theta \in \Theta\}} = \text{Var}\left(\frac{\partial \log p_\theta(X_0)}{\partial \theta}\right)$

Maximum de vraisemblance - idée de preuve

- $\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow \mathcal{N}(0, F^{-1}(\theta_*))$

On note $\hat{l}_n(\theta) = \frac{1}{n} \sum \log p_\theta(x_i)$, $\forall \theta$

$\hat{\theta}_n$ est choisi tel que $\hat{l}'_n(\hat{\theta}_n) = 0$

$\hat{\theta}_n$ proche de θ_* $\rightarrow \hat{l}'_n(\theta) \simeq \hat{l}'_n(\theta_*) + (\hat{\theta}_n - \theta_*)\hat{l}''_n(\theta_*)$

On a donc $\hat{l}'_n(\theta_*) + (\hat{\theta}_n - \theta_*)\hat{l}''_n(\theta_*) \simeq 0$ puis
 $\sqrt{n}\hat{l}'_n(\theta_*) + (\sqrt{n}(\hat{\theta}_n - \theta_*))\hat{l}''_n(\theta_*) \simeq 0$

Or, le théorème central limite donne $\sqrt{n}\hat{l}'_n(\theta_*) \rightarrow \mathcal{N}(0, F_{\theta_*})$

et la loi des grands nombres donne $\hat{l}''_n(\theta_*) \rightarrow -F_{\theta_*}$

On finit donc avec $\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow \mathcal{N}(0, F_{\theta_*}^{-1})$

Modèles de mélange et classification

Modèle de mélange et classification

Un individu peut être issu de deux catégories, A ou B .

Pour un individu x on peut mesurer une caractéristique $t(x)$

- Si $x \in A$, $t(x) \sim \mathcal{N}(\mu_A, \sigma_A^2)$
- Si $x \in B$, $t(x) \sim \mathcal{N}(\mu_B, \sigma_B^2)$

On suppose que les individus des différentes classes sont en proportions π_A et π_B ,
 $\pi_A + \pi_B = 1$.

Modèle de mélange et classification

Un individu peut être issu de deux catégories, A ou B .

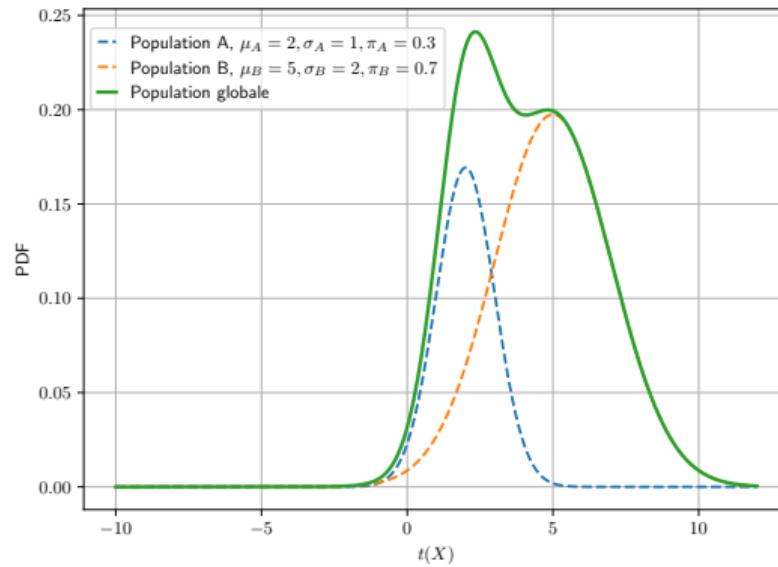
Pour un individu x on peut mesurer une caractéristique $t(x)$

- Si $x \in A$, $t(x) \sim \mathcal{N}(\mu_A, \sigma_A^2)$
- Si $x \in B$, $t(x) \sim \mathcal{N}(\mu_B, \sigma_B^2)$

On suppose que les individus des différentes classes sont en proportions π_A et π_B ,
 $\pi_A + \pi_B = 1$.

? On tire un individu X au hasard et on observe $t(x)$, quelle est la densité de probabilité p_θ de $t(X)$ et quels paramètres θ décrivent la manipulation ?

Modèle de mélange et classification



Modèle de mélange et classification

Pour un nouvel individu x dont la classe est inconnue, on mesure $t(x)$.
L'individu est-il de la catégorie A ou de la catégorie B ?

Modèle de mélange et classification

Pour un nouvel individu x dont la classe est inconnue, on mesure $t(x)$.
L'individu est-il de la catégorie A ou de la catégorie B ?

$$P(x \in A | t(x)) = \frac{P(t(x) | x \in A) \cdot \pi_A}{P(t(x))} \propto P(t(x) | x \in A) \cdot \pi_A \triangleq q(A)$$
$$P(x \in B | t(x)) = \frac{P(t(x) | x \in B) \cdot \pi_B}{P(t(x))} \propto P(t(x) | x \in B) \cdot \pi_B \triangleq q(B)$$

Modèle de mélange et classification

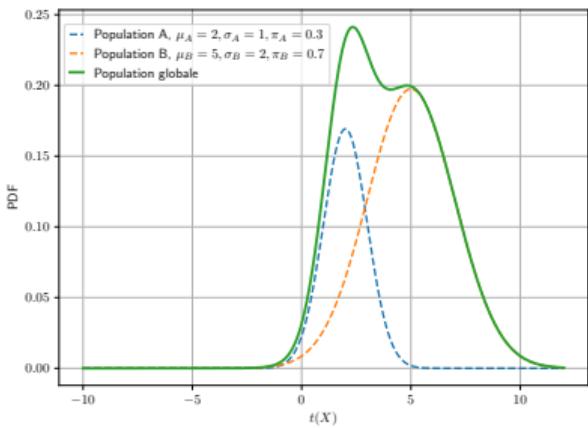
Pour un nouvel individu x dont la classe est inconnue, on mesure $t(x)$.
L'individu est-il de la catégorie A ou de la catégorie B ?

$$P(x \in A|t(x)) = \frac{P(t(x)|x \in A) \cdot \pi_A}{P(t(x))} \propto P(t(x)|x \in A) \cdot \pi_A \triangleq q(A)$$
$$P(x \in B|t(x)) = \frac{P(t(x)|x \in B) \cdot \pi_B}{P(t(x))} \propto P(t(x)|x \in B) \cdot \pi_B \triangleq q(B)$$

$$P(x \in A|t(x)) = \frac{q(A)}{q(A) + q(B)} \text{ et } P(x \in B|t(x)) = \frac{q(B)}{q(A) + q(B)}$$

Modèle de mélange et classification - Exemple numérique

On observe $t(x) = 3$



$$q(A) = P(t(x)|x \in A) \cdot \pi_A \propto \frac{1}{2} \exp -\frac{(3-2)^2}{2 \cdot 1} \cdot 0.3 \simeq 0.18$$

$$q(B) = P(t(x)|x \in B) \cdot \pi_B \propto \frac{1}{2} \exp -\frac{(3-5)^2}{2 \cdot 4} \cdot 0.7 \simeq 0.27$$

$$P(x \in A|t(x)) = \frac{q(A)}{q(A)+q(B)} = 0.4$$

$$P(x \in B|t(x)) = \frac{q(B)}{q(A)+q(B)} = 0.6$$

Modèle de mélange et classification, N classes

On généralise le problème à N classes A_i , pour chaque classe, la mesure t est distribuée selon une densité p_{θ_i} .

Les classes sont en proportions $\pi_i \geq 0$, $\sum \pi_i = 1$.

? On a un individu x dont la mesure vaut $t(x)$, de quelle classe est-il issu ?

Modèle de mélange et classification, N classes

On généralise le problème à N classes A_i , pour chaque classe, la mesure t est distribuée selon une densité p_{θ_i} .

Les classes sont en proportions $\pi_i \geq 0$, $\sum \pi_i = 1$.

? On a un individu x dont la mesure vaut $t(x)$, de quelle classe est-il issu ?

- Pour chaque classe on calcule $l_i \triangleq \log p_{\theta_i}(t(x)) + \log \pi_i$.

Modèle de mélange et classification, N classes

On généralise le problème à N classes A_i , pour chaque classe, la mesure t est distribuée selon une densité p_{θ_i} .

Les classes sont en proportions $\pi_i \geq 0$, $\sum \pi_i = 1$.

? On a un individu x dont la mesure vaut $t(x)$, de quelle classe est-il issu ?

- Pour chaque classe on calcule $l_i \triangleq \log p_{\theta_i}(t(x)) + \log \pi_i$.
- On note $p_i = P(x \in A_i | t(x))$

Modèle de mélange et classification, N classes

On généralise le problème à N classes A_i , pour chaque classe, la mesure t est distribuée selon une densité p_{θ_i} .

Les classes sont en proportions $\pi_i \geq 0$, $\sum \pi_i = 1$.

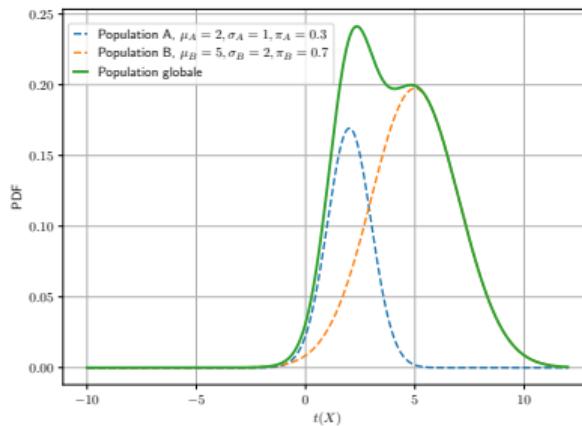
? On a un individu x dont la mesure vaut $t(x)$, de quelle classe est-il issu ?

- Pour chaque classe on calcule $l_i \triangleq \log p_{\theta_i}(t(x)) + \log \pi_i$.
- On note $p_i = P(x \in A_i | t(x))$

- On calcule le vecteur de probabilités $\begin{pmatrix} p_0 \\ \dots \\ p_{N-1} \end{pmatrix} = \text{softmax} \left(\begin{pmatrix} l_0 \\ \dots \\ l_{N-1} \end{pmatrix} \right)$

$$\text{softmax} \left(\begin{pmatrix} l_0 \\ \dots \\ l_{N-1} \end{pmatrix} \right) = \frac{1}{\sum e^{l_i}} \begin{pmatrix} e^{l_0} \\ \dots \\ e^{l_{N-1}} \end{pmatrix}$$

Modèle de mélange et classification, estimation des paramètres



? Comment estimer $\theta = (\mu_A, \mu_B, \sigma_A, \sigma_B, \pi_A, \pi_B)$ à partir de N observations $(x_i, t(x_i))$?

⇒ A voir en exercice

Modèles Markoviens

⇒ Modèles de variables aléatoires ordonnées non-indépendantes.

⇒ Modèles de variables aléatoires ordonnées non-indépendantes.

Une suite de variables aléatoires (Z_t) est une *chaîne de Markov* dans un ensemble fini $\mathcal{E} = \{e_0, \dots, e_{n-1}\}$, de loi initiale π_0 et de *noyau de transition* Π si:

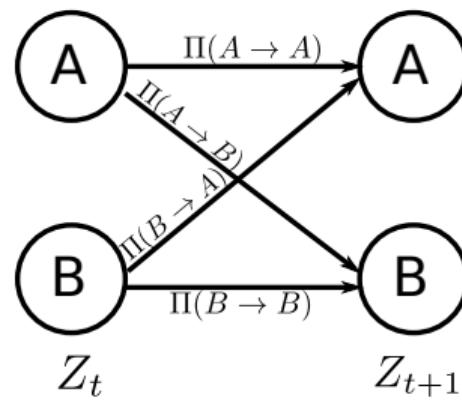
Chaînes de Markov

⇒ Modèles de variables aléatoires ordonnées non-indépendantes.

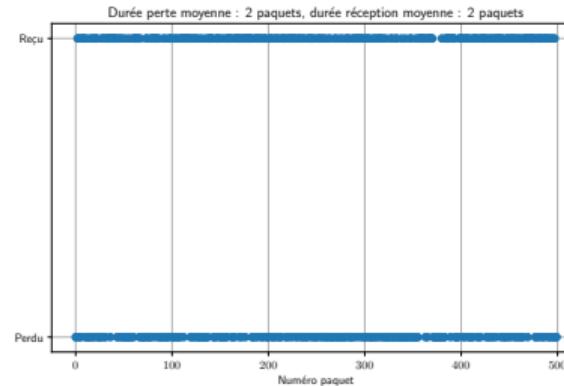
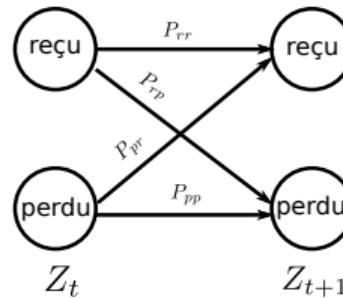
Une suite de variables aléatoires (Z_t) est une *chaîne de Markov* dans un ensemble fini $\mathcal{E} = \{e_0, \dots, e_{n-1}\}$, de loi initiale π_0 et de *noyau de transition* Π si:

- $Z_0 \sim \pi_0$
- Pour tout $y, z \in \mathcal{E}^2$, $\Pi(y \rightarrow z)$ est la probabilité de transition de y vers z :

$$\forall t, P(Z_{t+1} = z | Z_t = y) = \Pi(y \rightarrow z)$$

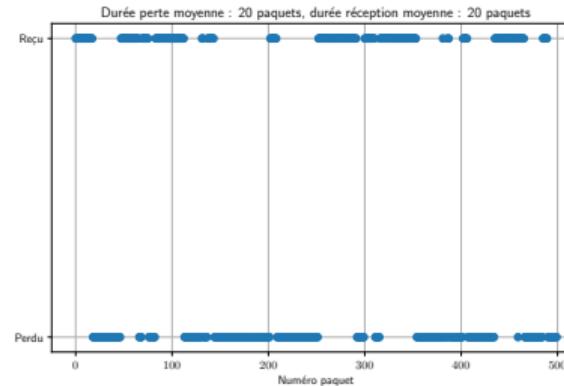
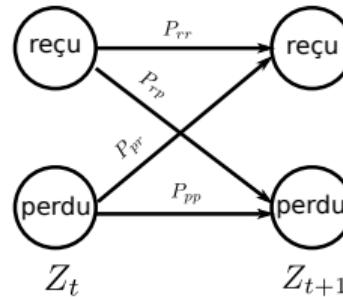


Chaînes de Markov - Exemple : Perte de paquets



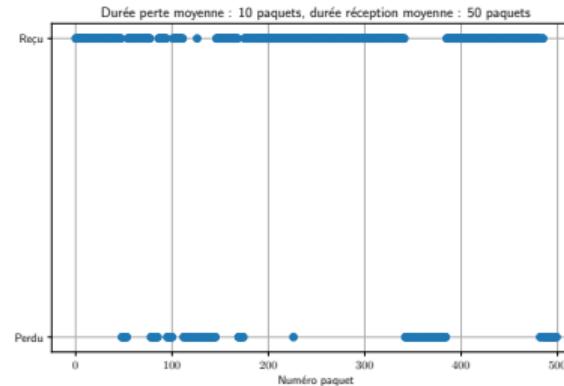
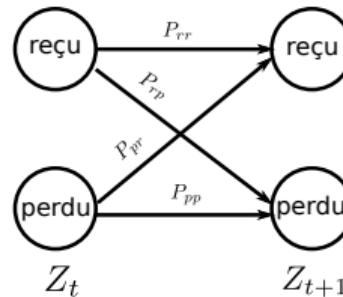
- Durée moyenne des zones perdues : $E_p = \frac{1}{P_{pr}}$ (cf espérance loi géométrique)
- Durée moyenne des zones bien reçues : $E_r = \frac{1}{P_{rp}}$ (idem)
- Proportion globale de paquets perdus : $P_p = \frac{E_p}{E_r+E_p}$
- Proportion globale de paquets bien reçus : $P_r = \frac{E_r}{E_r+E_p}$

Chaînes de Markov - Exemple : Perte de paquets



- Durée moyenne des zones perdues : $E_p = \frac{1}{P_{pr}}$ (cf espérance loi géométrique)
- Durée moyenne des zones bien reçues : $E_r = \frac{1}{P_{rp}}$ (idem)
- Proportion globale de paquets perdus : $P_p = \frac{E_p}{E_r+E_p}$
- Proportion globale de paquets bien reçus : $P_r = \frac{E_r}{E_r+E_p}$

Chaînes de Markov - Exemple : Perte de paquets



- Durée moyenne des zones perdues : $E_p = \frac{1}{P_{pr}}$ (cf espérance loi géométrique)
- Durée moyenne des zones bien reçues : $E_r = \frac{1}{P_{rp}}$ (idem)
- Proportion globale de paquets perdus : $P_p = \frac{E_p}{E_r+E_p}$
- Proportion globale de paquets bien reçus : $P_r = \frac{E_r}{E_r+E_p}$

Chaînes de Markov - Propriétés

On a la propriété de *mémoire courte*:

$$P(Z_{t+1}, Z_{t+2}, \dots | Z_0, \dots, Z_{t-1}, Z_t) = P(Z_{t+1}, Z_{t+2}, \dots | Z_t) \forall t$$

qu'on note aussi

$$P(Z_{t+1:\infty} | Z_{0:t}) = P(Z_{t+1:\infty} | Z_t) \forall t$$

Chaînes de Markov - Propriétés

On a la propriété de *mémoire courte*:

$$P(Z_{t+1}, Z_{t+2}, \dots | Z_0, \dots, Z_{t-1}, Z_t) = P(Z_{t+1}, Z_{t+2}, \dots | Z_t) \forall t$$

qu'on note aussi

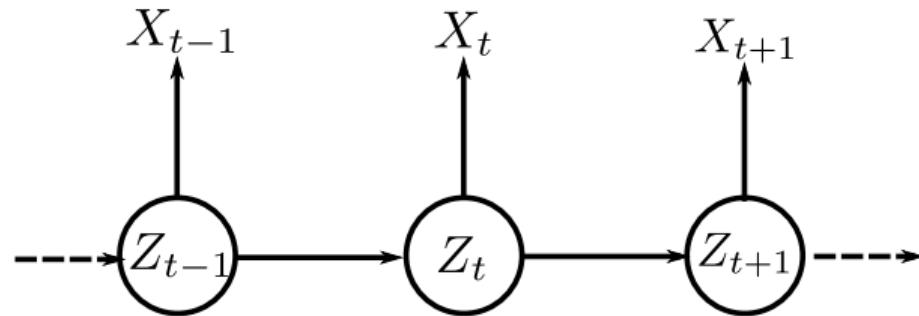
$$P(Z_{t+1:\infty} | Z_{0:t}) = P(Z_{t+1:\infty} | Z_t) \forall t$$

Et la propriété *d'indépendance conditionnelle*:

$$P(Z_{t-k}, Z_{t+l} | Z_t) = P(Z_{t-k} | Z_t) P(Z_{t+l} | Z_t) \forall t, k, l > 0, k \leq t$$

Modèles de Markov cachés

Un *modèle de Markov caché* est une chaîne de Markov à laquelle on ajoute pour chaque temps t une variable dite de diffusion X_t .
La loi de X_t est définie conditionnellement à Z_t : $P(X_t|Z_t)$.

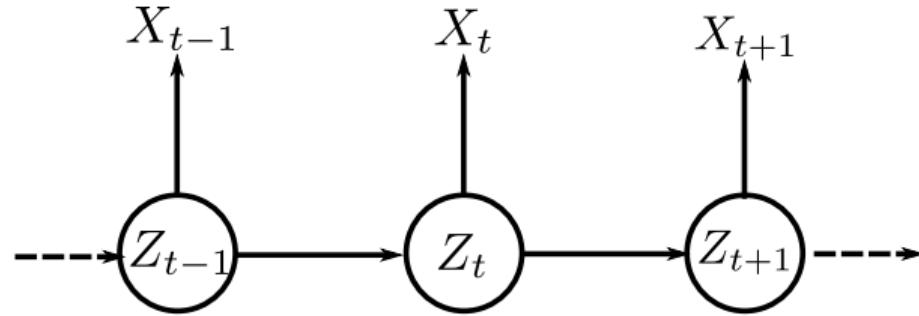


Modèles de Markov cachés - Propriétés

On a encore une propriété d'indépendance conditionnelle:

$$P(X_{t-k}, X_{t+k} | Z_t) = P(X_{t-k} | Z_t)P(X_{t+k} | Z_t) \quad \forall t, k, l > 0, k \leq t$$

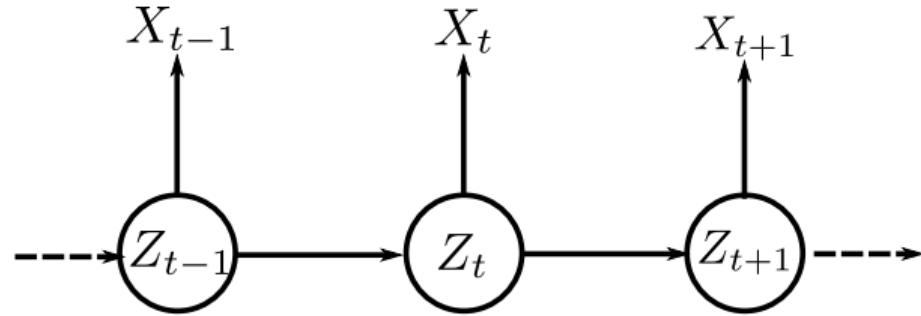
Modèles de Markov cachés - Problèmes d'inférence



On a deux problèmes d'inférence sur les modèles de Markov cachés:

- Estimer $P(Z_t = z|x_{0:T-1})$, $\forall z \in \mathcal{E}$, $\forall t < T$

Modèles de Markov cachés - Problèmes d'inférence



On a deux problèmes d'inférence sur les modèles de Markov cachés:

- Estimer $P(Z_t = z|x_{0:T-1}), \forall z \in \mathcal{E}, \forall t < T$
- Estimer la séquence $(z_{0:T-1}^*) = \text{argmax}(P(z_{0:T-1}|x_{0:T-1}))$

Algorithme Froward Backward

Estimer $P(Z_t = z | x_{0:T-1})$, $\forall z \in \mathcal{E}$

$$\gamma_t(z) = P(Z_t = z | X_{0:T-1} = x_{0:T-1}) = \frac{P(Z_t = z \cap X_{0:T-1} = x_{0:T-1})}{P(X_{0:T-1} = x_{0:T-1})}$$

Algorithme Froward Backward

Estimer $P(Z_t = z | x_{0:T-1})$, $\forall z \in \mathcal{E}$

$$\gamma_t(z) = P(Z_t = z | X_{0:T-1} = x_{0:T-1}) = \frac{P(Z_t = z \cap X_{0:T-1} = x_{0:T-1})}{P(X_{0:T-1} = x_{0:T-1})}$$

$$\gamma_t(z) = \frac{1}{\Omega} P(Z_t = z \cap X_{0:t} = x_{0:t}) \cdot P(X_{t+1:T-1} = x_{t+1:T-1} | Z_t = z \cap X_{0:t} = x_{0:t})$$

Algorithme Froward Backward

Estimer $P(Z_t = z | x_{0:T-1})$, $\forall z \in \mathcal{E}$

$$\gamma_t(z) = P(Z_t = z | X_{0:T-1} = x_{0:T-1}) = \frac{P(Z_t = z \cap X_{0:T-1} = x_{0:T-1})}{P(X_{0:T-1} = x_{0:T-1})}$$

$$\gamma_t(z) = \frac{1}{\Omega} P(Z_t = z \cap X_{0:t} = x_{0:t}) \cdot P(X_{t+1:T-1} = x_{t+1:T-1} | Z_t = z \cap X_{0:t} = x_{0:t})$$

$$\gamma_t(z) = \underbrace{\frac{1}{\Omega} P(Z_t = z \cap X_{0:t} = x_{0:t})}_{f_{0:t}(z)} \cdot \underbrace{P(X_{t+1:T-1} = x_{t+1:T-1} | Z_t = z)}_{b_{t:T-1}(z)}$$

Algorithme Froward Backward

Estimer $P(Z_t = z | x_{0:T-1})$, $\forall z \in \mathcal{E}$

$$\gamma_t(z) = P(Z_t = z | X_{0:T-1} = x_{0:T-1}) = \frac{P(Z_t = z \cap X_{0:T-1} = x_{0:T-1})}{P(X_{0:T-1} = x_{0:T-1})}$$

$$\gamma_t(z) = \frac{1}{\Omega} P(Z_t = z \cap X_{0:t} = x_{0:t}) \cdot P(X_{t+1:T-1} = x_{t+1:T-1} | Z_t = z \cap X_{0:t} = x_{0:t})$$

$$\gamma_t(z) = \underbrace{\frac{1}{\Omega} P(Z_t = z \cap X_{0:t} = x_{0:t})}_{f_{0:t}(z)} \cdot \underbrace{P(X_{t+1:T-1} = x_{t+1:T-1} | Z_t = z)}_{b_{t:T-1}(z)}$$

- $f_{0:t}(z)$ se calcule par récurrence:

$$\forall z, f_{0:t}(z) = \sum_y f_{0:t-1}(y) \cdot \Pi(y \rightarrow z) \cdot P(X_t = x_t | Z_t = z)$$

Algorithme Froward Backward

Estimer $P(Z_t = z | x_{0:T-1})$, $\forall z \in \mathcal{E}$

$$\gamma_t(z) = P(Z_t = z | X_{0:T-1} = x_{0:T-1}) = \frac{P(Z_t = z \cap X_{0:T-1} = x_{0:T-1})}{P(X_{0:T-1} = x_{0:T-1})}$$

$$\gamma_t(z) = \frac{1}{\Omega} P(Z_t = z \cap X_{0:t} = x_{0:t}) \cdot P(X_{t+1:T-1} = x_{t+1:T-1} | Z_t = z \cap X_{0:t} = x_{0:t})$$

$$\gamma_t(z) = \underbrace{\frac{1}{\Omega} P(Z_t = z \cap X_{0:t} = x_{0:t})}_{f_{0:t}(z)} \cdot \underbrace{P(X_{t+1:T-1} = x_{t+1:T-1} | Z_t = z)}_{b_{t:T-1}(z)}$$

- $f_{0:t}(z)$ se calcule par récurrence:

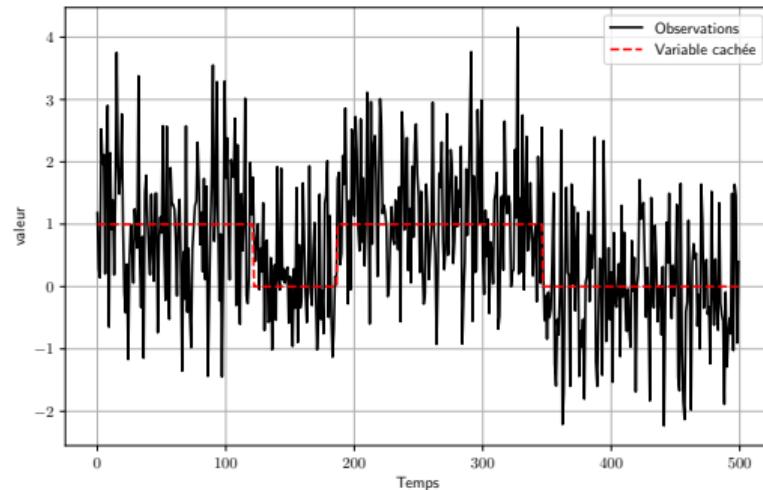
$$\forall z, f_{0:t}(z) = \sum_y f_{0:t-1}(y) \cdot \Pi(y \rightarrow z) \cdot P(X_t = x_t | Z_t = z)$$

- $b_{t:T-1}(z)$ se calcule par récurrence aussi:

$$\forall z, b_{t:T-1}(z) = \sum_y b_{t+1:T-1}(y) \cdot \Pi(z \rightarrow y) \cdot P(X_{t+1} = x_{t+1} | Z_{t+1} = y)$$

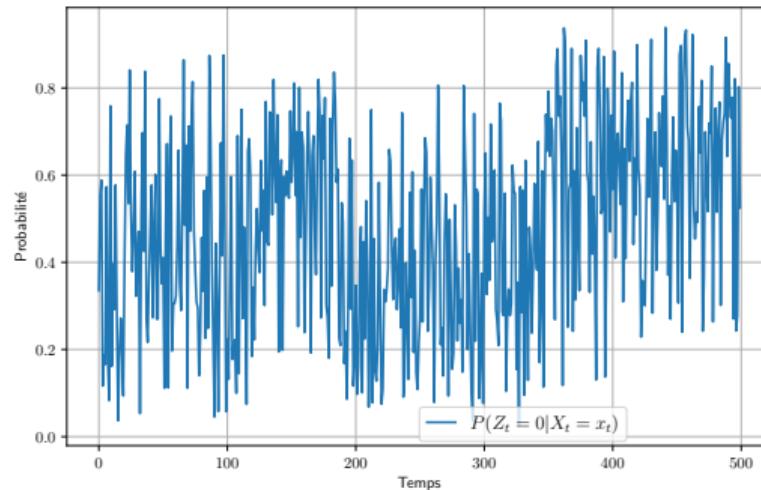
Algorithme Foward Backward - Exemple

Deux états, A et B , $\Pi(A \rightarrow A) = \Pi(B \rightarrow B) = 0.99$, $(Y_t | Z_t = A) \sim \mathcal{N}(0, 1)$,
 $(Y_t | Z_t = B) \sim \mathcal{N}(1, 1)$



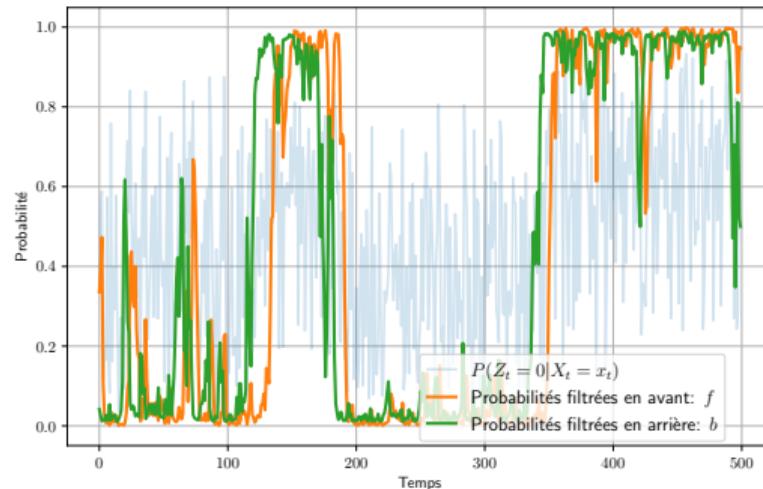
Algorithme Froward Backward - Exemple

Deux états, A et B , $\Pi(A \rightarrow A) = \Pi(B \rightarrow B) = 0.99$, $(Y_t | Z_t = A) \sim \mathcal{N}(0, 1)$,
 $(Y_t | Z_t = B) \sim \mathcal{N}(1, 1)$



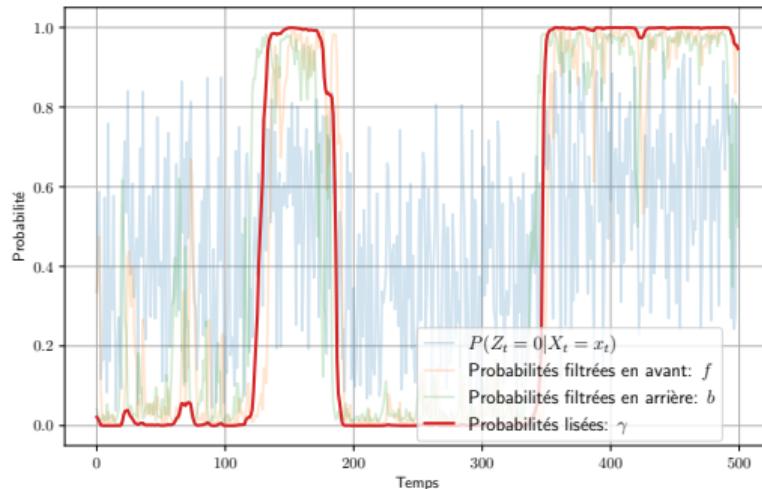
Algorithme Foward Backward - Exemple

Deux états, A et B , $\Pi(A \rightarrow A) = \Pi(B \rightarrow B) = 0.99$, $(Y_t | Z_t = A) \sim \mathcal{N}(0, 1)$,
 $(Y_t | Z_t = B) \sim \mathcal{N}(1, 1)$



Algorithme Froward Backward - Exemple

Deux états, A et B , $\Pi(A \rightarrow A) = \Pi(B \rightarrow B) = 0.99$, $(Y_t | Z_t = A) \sim \mathcal{N}(0, 1)$, $(Y_t | Z_t = B) \sim \mathcal{N}(1, 1)$



Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$

Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$

On a $P(z_{0:T-1} | x_{0:T-1}) \propto P(z_{0:T-1} \cap x_{0:T-1})$

Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$

On a $P(z_{0:T-1} | x_{0:T-1}) \propto P(z_{0:T-1} \cap x_{0:T-1})$

On pose pour tout $z \in \mathcal{E}$:

- $\Phi_t(z, z_{0:t-1}) = P(Z_t = z, X_t = x_t, Z_{0:t-1} = z_{0:t-1}, X_{0:t-1} = x_{0:t-1})$
- $\Psi_t(z) = \max_{z_{0:t-1}} \Phi_t(z, z_{0:t-1}) \quad \delta_t(z) = \operatorname{argmax}_{z_{0:t-1}} \Phi_t(z, z_{0:t-1})$

Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$

On a $P(z_{0:T-1} | x_{0:T-1}) \propto P(z_{0:T-1} \cap x_{0:T-1})$

On pose pour tout $z \in \mathcal{E}$:

- $\Phi_t(z, z_{0:t-1}) = P(Z_t = z, X_t = x_t, Z_{0:t-1} = z_{0:t-1}, X_{0:t-1} = x_{0:t-1})$
- $\Psi_t(z) = \max_{z_{0:t-1}} \Phi_t(z, z_{0:t-1}) \quad \delta_t(z) = \operatorname{argmax}_{z_{0:t-1}} \Phi_t(z, z_{0:t-1})$

Relation de récurrence en notant y la valeur prise par Z_{t-1} :

$$\Phi_t(z, z_{0:t-1}) = P(Z_t = z, X_t = x_t | Z_{0:t-1} = z_{0:t-1}, X_{0:t-1} = x_{0:t-1}) \cdot P(Z_{0:t-1} = z_{0:t-1}, X_{0:t-1} = x_{0:t-1})$$

$$\Phi_t(z, z_{0:t-1}) = P(Z_t = z, X_t = x_t | Z_{t-1} = y) \cdot \Phi_{t-1}(y, z_{0:t-2})$$

$$\Phi_t(z, z_{0:t-1}) = P(X_t = x_t | Z_t = z) \cdot P(Z_t = z | Z_{t-1} = y) \cdot \Phi_{t-1}(y, z_{0:t-2})$$

Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$

On a $P(z_{0:T-1} | x_{0:T-1}) \propto P(z_{0:T-1} \cap x_{0:T-1})$

On pose pour tout $z \in \mathcal{E}$:

- $\Phi_t(z, z_{0:t-1}) = P(Z_t = z, X_t = x_t, Z_{0:t-1} = z_{0:t-1}, X_{0:t-1} = x_{0:t-1})$
- $\Psi_t(z) = \max_{z_{0:t-1}} \Phi_t(z, z_{0:t-1}) \quad \delta_t(z) = \operatorname{argmax}_{z_{0:t-1}} \Phi_t(z, z_{0:t-1})$

Relation de récurrence en notant y la valeur prise par Z_{t-1} :

$$\Phi_t(z, z_{0:t-1}) = P(Z_t = z, X_t = x_t | Z_{0:t-1} = z_{0:t-1}, X_{0:t-1} = x_{0:t-1}) \cdot P(Z_{0:t-1} = z_{0:t-1}, X_{0:t-1} = x_{0:t-1})$$

$$\Phi_t(z, z_{0:t-1}) = P(Z_t = z, X_t = x_t | Z_{t-1} = y) \cdot \Phi_{t-1}(y, z_{0:t-2})$$

$$\Phi_t(z, z_{0:t-1}) = P(X_t = x_t | Z_t = z) \cdot P(Z_t = z | Z_{t-1} = y) \cdot \Phi_{t-1}(y, z_{0:t-2})$$

$$\Psi_t(z) = \max_y P(X_t = x_t | Z_t = z) \cdot \Pi(y \rightarrow z) \cdot \Psi_{t-1}(y)$$

Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$
⇒ Programmation dynamique

Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$

⇒ Programmation dynamique

- Initialisation

- On prend $\Psi_0(z) = P(Z_0 = z, X_0 = x_0) = P(X_0 = x_0 | Z_0 = z_0) \cdot P(Z_0 = z_0), \forall z$

Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$

\Rightarrow Programmation dynamique

- Initialisation

- On prend $\Psi_0(z) = P(Z_0 = z, X_0 = x_0) = P(X_0 = x_0 | Z_0 = z_0) \cdot P(Z_0 = z_0), \forall z$

- Récursion avant

- Pour $t = 1, \dots, T - 1,$

- $\Psi_t(z) = \max_y \Psi_{t-1}(y) \Pi(y \rightarrow z) P(X_t = x_t | Z_t = z)$

- $\delta_t(z) = \operatorname{argmax}_y \Psi_{t-1}(y) \Pi(y \rightarrow z) P(X_t = x_t | Z_t = z)$

Algorithme de Viterbi

Estimer $(z_0^* \cdots z_{T-1}^*) = \operatorname{argmax}_{z_{0:T-1}} P(z_{0:T-1} | x_{0:T-1})$ en observant $x_{0:T-1}$
⇒ Programmation dynamique

- Initialisation

- On prend $\Psi_0(z) = P(Z_0 = z, X_0 = x_0) = P(X_0 = x_0 | Z_0 = z_0) \cdot P(Z_0 = z_0), \forall z$

- Récursion avant

- Pour $t = 1, \dots, T - 1,$

- $\Psi_t(z) = \max_y \Psi_{t-1}(y) \Pi(y \rightarrow z) P(X_t = x_t | Z_t = z)$
 - $\delta_t(z) = \operatorname{argmax}_y \Psi_{t-1}(y) \Pi(y \rightarrow z) P(X_t = x_t | Z_t = z)$

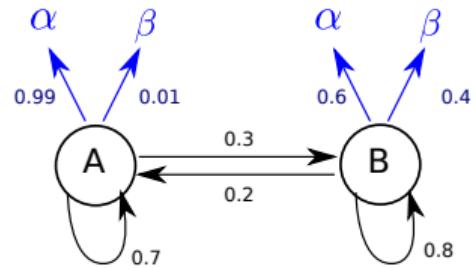
- Récursion arrière

- $z_{T-1}^* = \operatorname{argmax} \Psi_{T-1}$

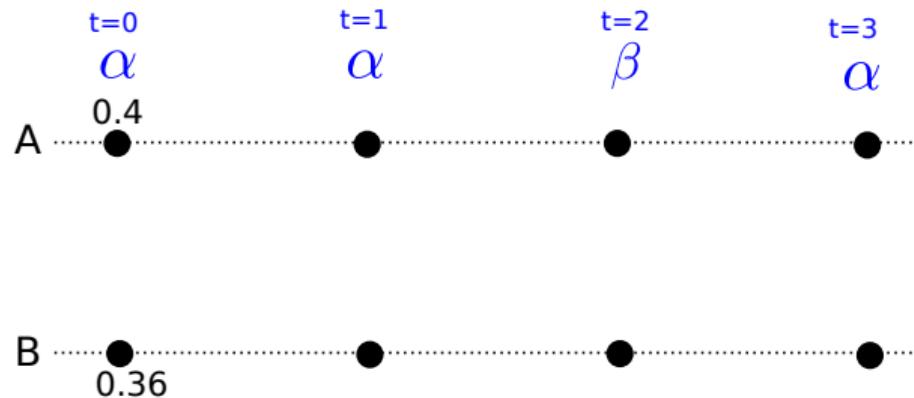
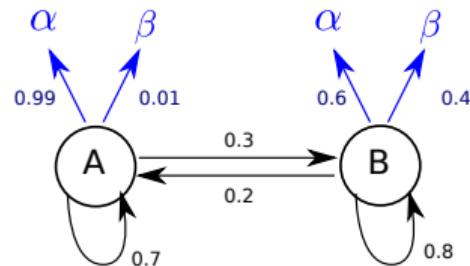
- Pour $t = T - 2, \dots, 0,$

- $z_t^* = \delta_{t+1}(z_{t+1}^*)$

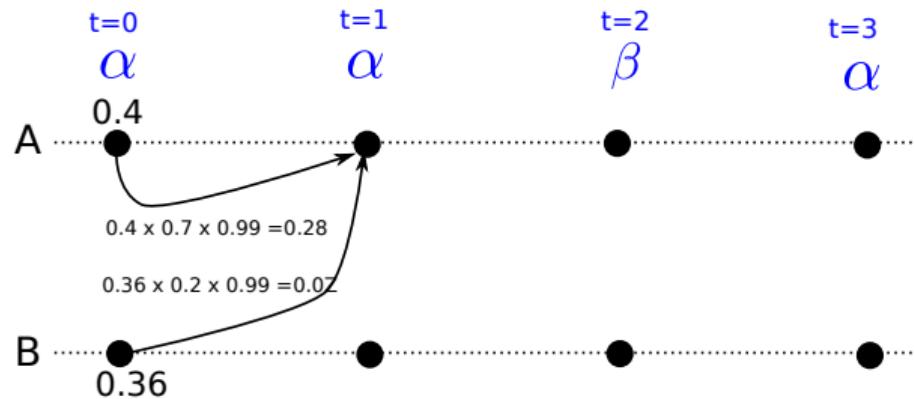
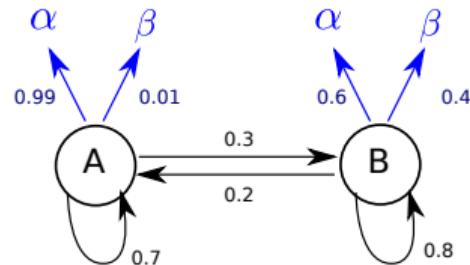
Algorithme de Viterbi



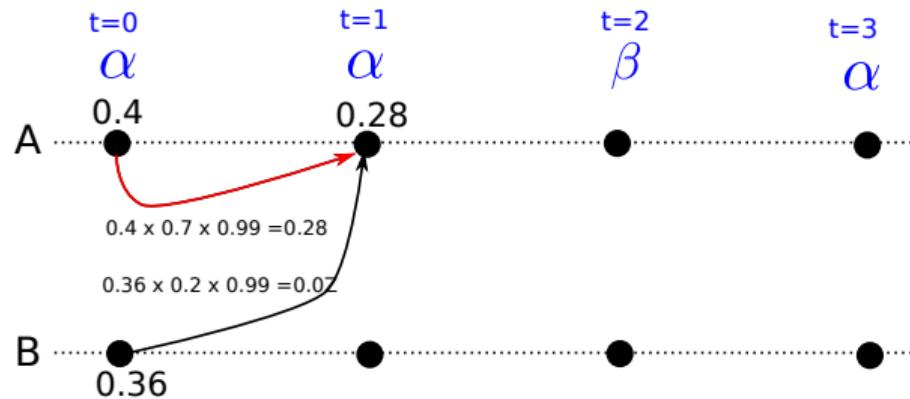
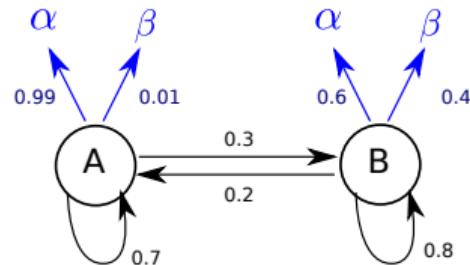
Algorithme de Viterbi



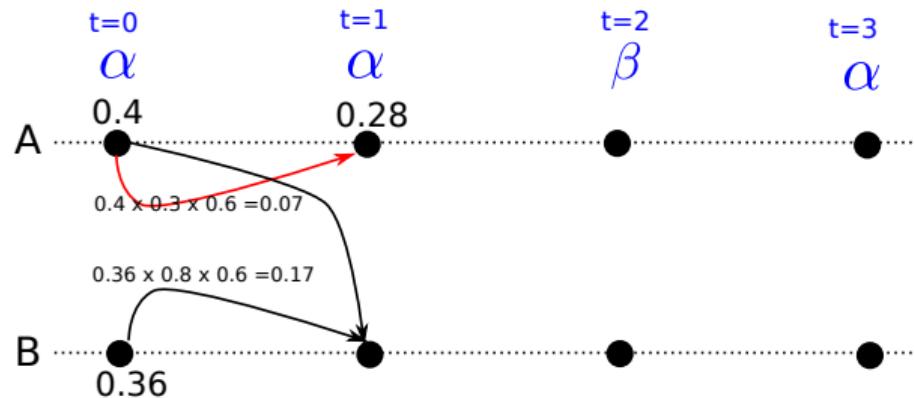
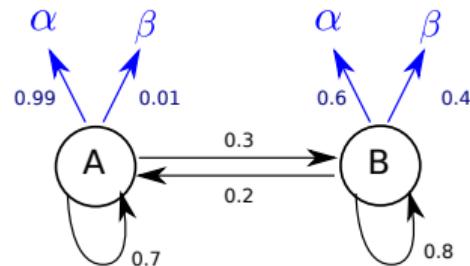
Algorithme de Viterbi



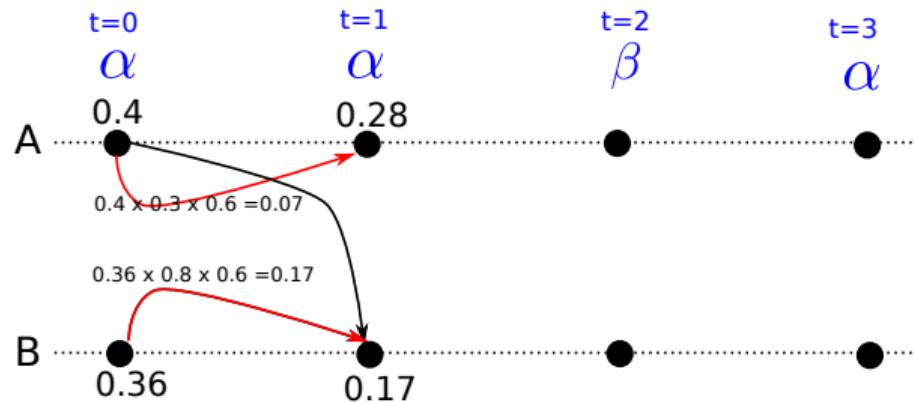
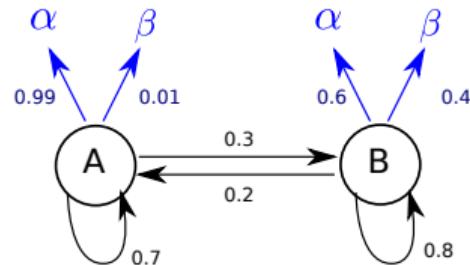
Algorithme de Viterbi



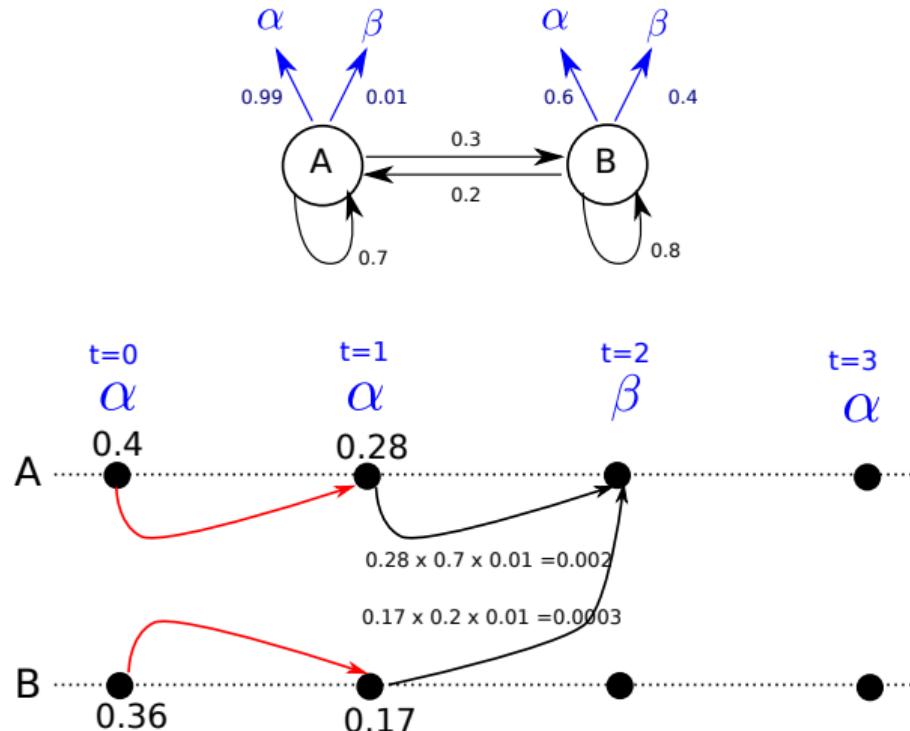
Algorithme de Viterbi



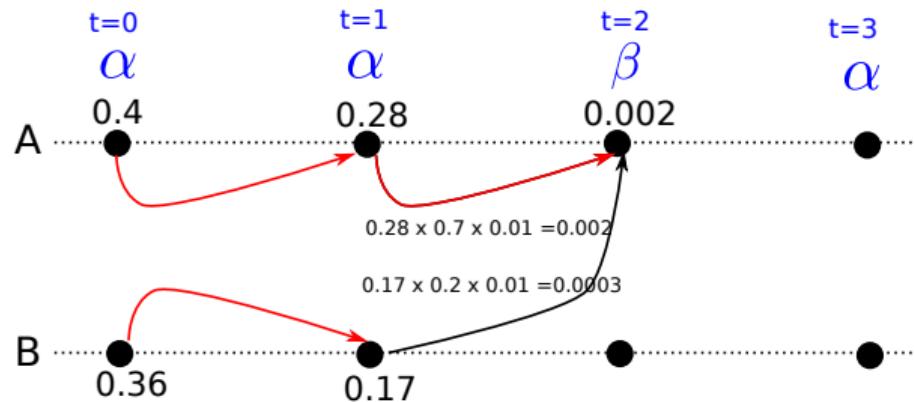
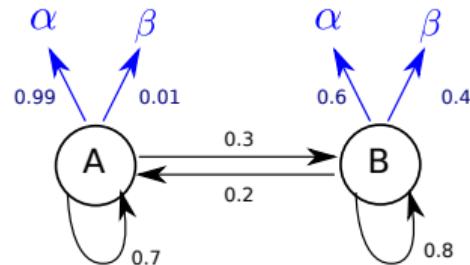
Algorithme de Viterbi



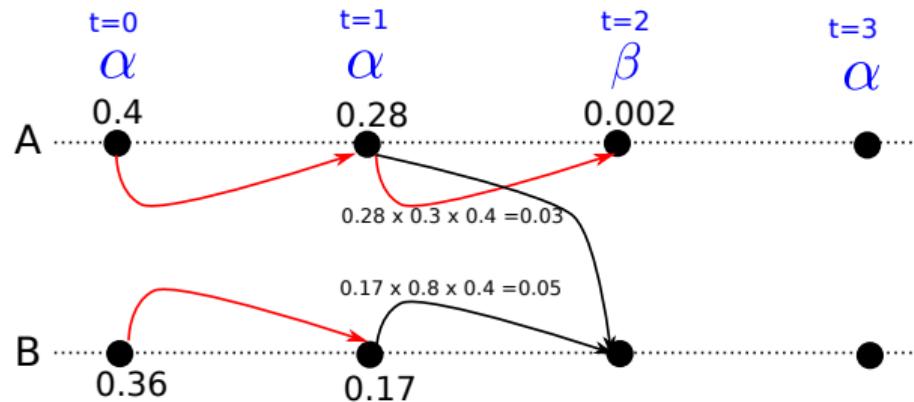
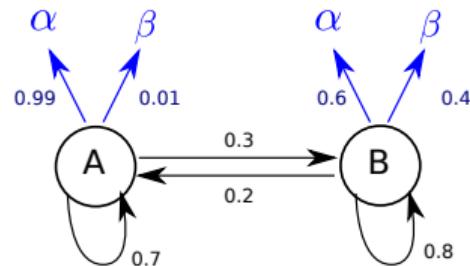
Algorithme de Viterbi



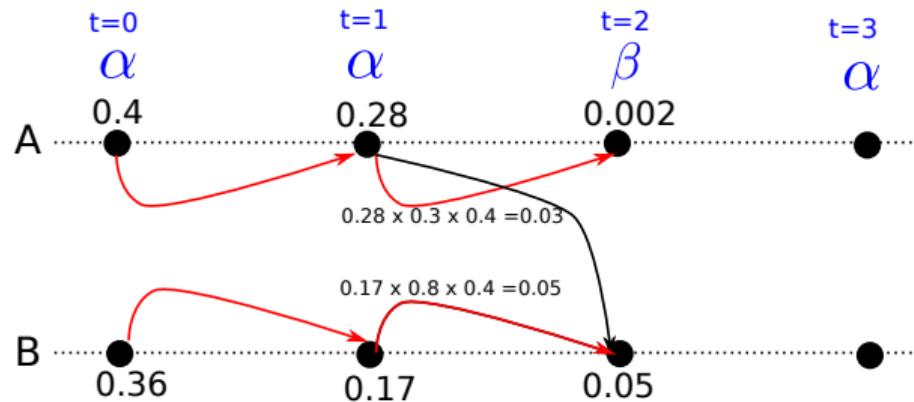
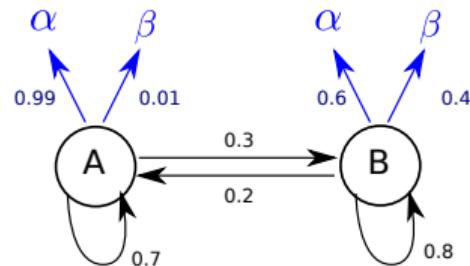
Algorithme de Viterbi



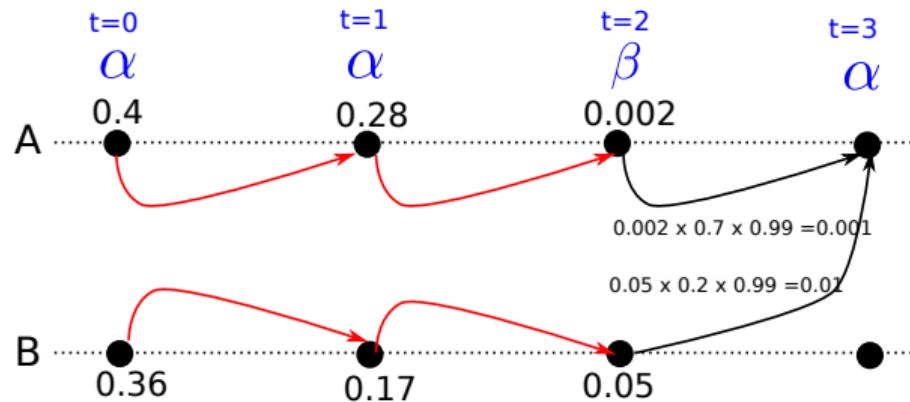
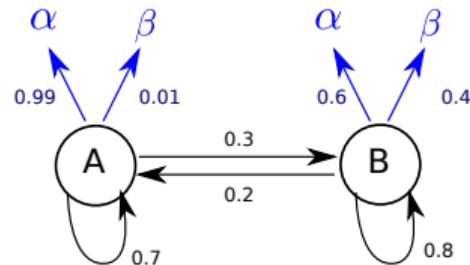
Algorithme de Viterbi



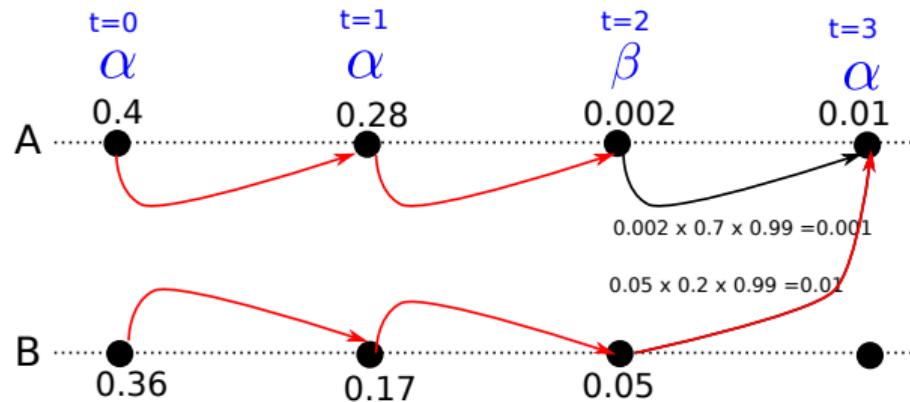
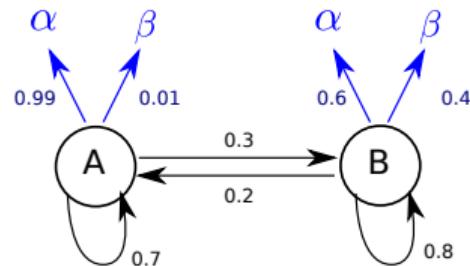
Algorithme de Viterbi



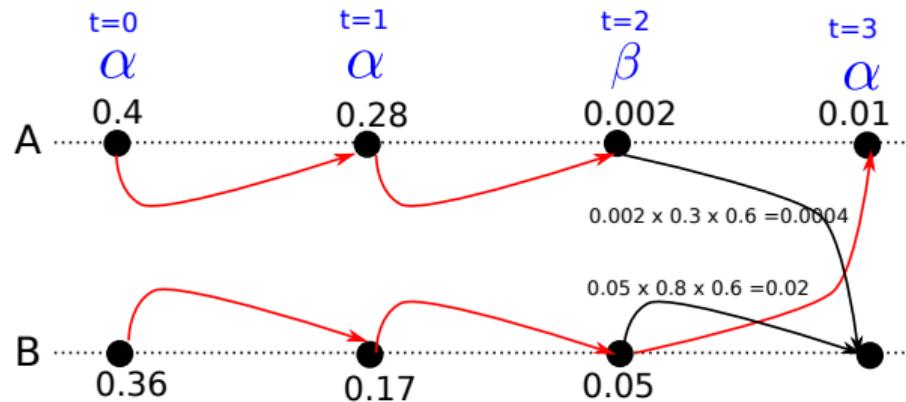
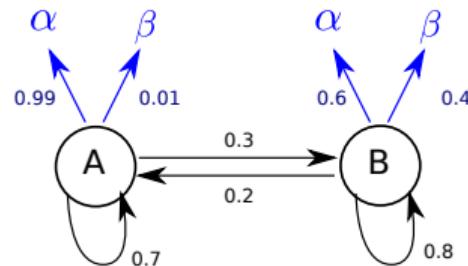
Algorithme de Viterbi



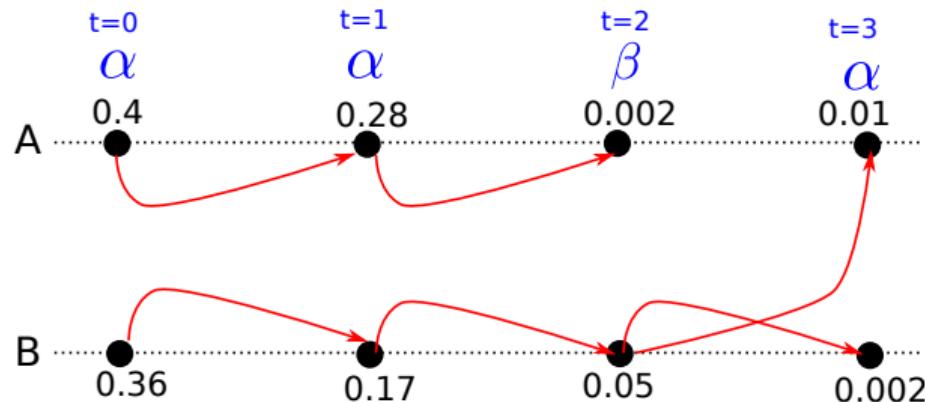
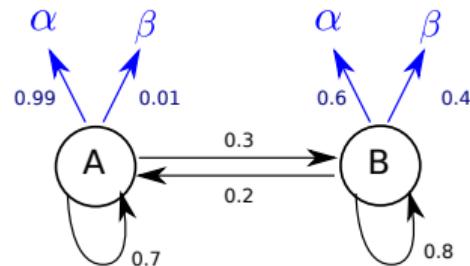
Algorithme de Viterbi



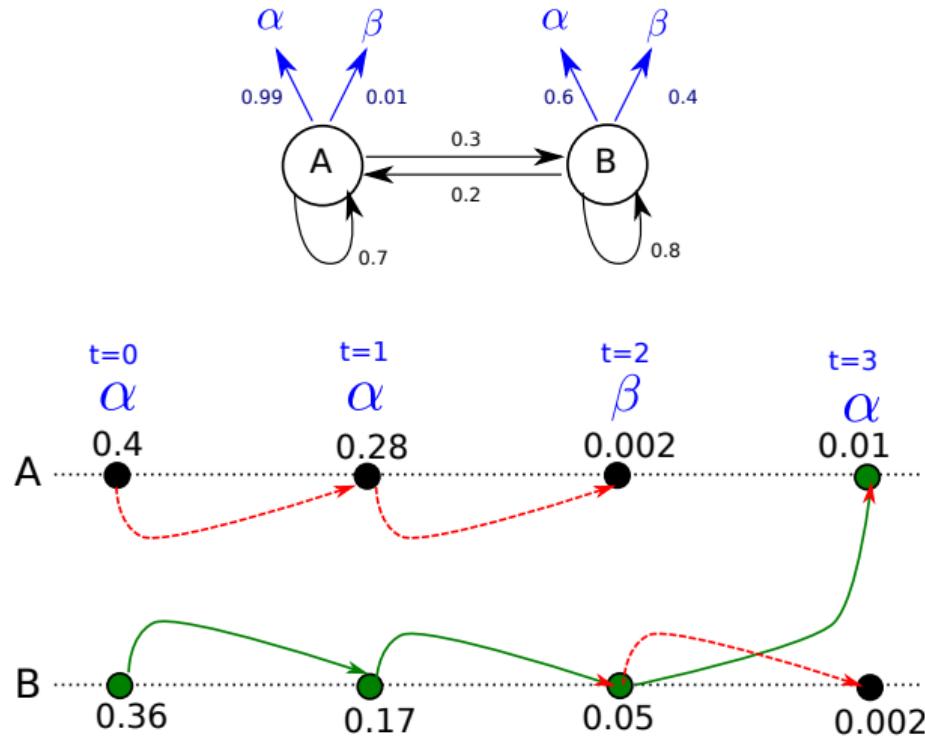
Algorithme de Viterbi



Algorithme de Viterbi



Algorithme de Viterbi



Estimation de paramètres pour les modèles de Markov

Il est facile d'estimer les paramètres d'un chaîne de Markov / d'un processus de Markov caché à partir d'un ensemble de trajectoires $\{((x_t^{(i)}, z_t^{(i)}))_t\}$.

Estimation de paramètres pour les modèles de Markov

Il est facile d'estimer les paramètres d'un chaîne de Markov / d'un processus de Markov caché à partir d'un ensemble de trajectoires $\{((x_t^{(i)}, z_t^{(i)}))_t\}$.

?

Le vérifier en exercice

Application des modèles Markoviens

- En télécom pour définir et décoder les codes correcteurs
- En reconnaissance de la parole pour prendre en compte la persistance des paramétrés vocaux
- En traitement du langage pour assurer la cohérence à l'échelle d'une phrase
- D'une manière générale, pour exploiter la corrélation d'événements dans le temps et fiabiliser les prédictions

Inférence Bayésienne

Un mot sur l'inférence Bayésienne

En *inférence Bayésienne*, on voit les paramètres du modèle statistique sous-jacent à une expérience comme des variables aléatoires:

$$p_{\theta}(X = x) \rightarrow P(X = x | \Theta = \theta)$$

Un mot sur l'inférence Bayésienne

En *inférence Bayésienne*, on voit les paramètres du modèle statistique sous-jacent à une expérience comme des variables aléatoires:

$$p_{\theta}(X = x) \rightarrow P(X = x | \Theta = \theta)$$

Puis on définit la loi de θ conditionnellement aux données observées via:

$$P(\Theta = \theta | X = x) = \frac{P(X = x | \Theta = \theta) \cdot P(\Theta = \theta)}{P(X = x)}$$

Inférence Bayésienne - Exemple 1, tirage aléatoire

Dans une urne on a n_A boules de types A et n_B boules de type B.

On note $p_A = \frac{n_A}{n_A + n_B}$ et $p_B = \frac{n_B}{n_A + n_B}$

Inférence Bayésienne - Exemple 1, tirage aléatoire

Dans une urne on a n_A boules de types A et n_B boules de type B.

On note $p_A = \frac{n_A}{n_A + n_B}$ et $p_B = \frac{n_B}{n_A + n_B}$

? On tire une boule, elle est de type A: $X_0 = A$, quelle est la distribution $P(p_A = p | X_0 = A)$?

Inférence Bayésienne - Exemple 1, tirage aléatoire

Dans une urne on a n_A boules de types A et n_B boules de type B.

On note $p_A = \frac{n_A}{n_A + n_B}$ et $p_B = \frac{n_B}{n_A + n_B}$

? On tire une boule, elle est de type A: $X_0 = A$, quelle est la distribution $P(p_A = p | X_0 = A)$?

$$P(p_A = p | X_0 = A) = \frac{P(X_0 = A | p_A = p) \cdot P(p_A = p)}{P(X_0 = A)}$$

Inférence Bayésienne - Exemple 1, tirage aléatoire

Dans une urne on a n_A boules de types A et n_B boules de type B.

On note $p_A = \frac{n_A}{n_A + n_B}$ et $p_B = \frac{n_B}{n_A + n_B}$

? On tire une boule, elle est de type A: $X_0 = A$, quelle est la distribution $P(p_A = p | X_0 = A)$?

$$P(p_A = p | X_0 = A) = \frac{P(X_0 = A | p_A = p) \cdot P(p_A = p)}{P(X_0 = A)}$$

$$P(p_A = p | X_0 = A) = \frac{p \cdot \mathbf{1}_{[0,1]}(p)}{P(X_0 = A)}$$

Inférence Bayésienne - Exemple 1, tirage aléatoire

Dans une urne on a n_A boules de types A et n_B boules de type B.

On note $p_A = \frac{n_A}{n_A + n_B}$ et $p_B = \frac{n_B}{n_A + n_B}$

? On tire une boule, elle est de type A: $X_0 = A$, quelle est la distribution $P(p_A = p | X_0 = A)$?

$$P(p_A = p | X_0 = A) = \frac{P(X_0 = A | p_A = p) \cdot P(p_A = p)}{P(X_0 = A)}$$

$$P(p_A = p | X_0 = A) = \frac{p \cdot \mathbf{1}_{[0,1]}(p)}{P(X_0 = A)}$$

$$Z = P(X_0 = A) = \text{constante de normalisation}, Z = \int_0^1 pdp = \frac{1}{2}$$

Inférence Bayésienne - Exemple 1, tirage aléatoire

Dans une urne on a n_A boules de types A et n_B boules de type B.

On note $p_A = \frac{n_A}{n_A + n_B}$ et $p_B = \frac{n_B}{n_A + n_B}$

? On tire une boule, elle est de type A: $X_0 = A$, quelle est la distribution $P(p_A = p | X_0 = A)$?

$$P(p_A = p | X_0 = A) = \frac{P(X_0 = A | p_A = p) \cdot P(p_A = p)}{P(X_0 = A)}$$

$$P(p_A = p | X_0 = A) = \frac{p \cdot \mathbf{1}_{[0,1]}(p)}{P(X_0 = A)}$$

$$Z = P(X_0 = A) = \text{constante de normalisation}, Z = \int_0^1 pdp = \frac{1}{2}$$

$$P(p_A = p | X_0 = A) = \frac{p}{2} \mathbf{1}_{[0,1]}(p)$$

Inférence Bayésienne - Exemple 1, tirage aléatoire

? On tire M boules, m_A de type A et m_B de type B quelle est la distribution $P(p_a = p | X_0, \dots, X_M)$?

Inférence Bayésienne - Exemple 1, tirage aléatoire

? On tire M boules, m_A de type A et m_B de type B quelle est la distribution $P(p_a = p | X_0, \dots, X_M)$?

$$P(p_A = p | X_0, \dots, X_M) = \frac{P(X_0, \dots, X_M | p_A = p) \cdot P(p_A = p)}{P(X_0, \dots, X_M)}$$

Inférence Bayésienne - Exemple 1, tirage aléatoire

? On tire M boules, m_A de type A et m_B de type B quelle est la distribution $P(p_a = p|X_0, \dots, X_M)$?

$$P(p_A = p|X_0, \dots, X_M) = \frac{P(X_0, \dots, X_M|p_A = p) \cdot P(p_A = p)}{P(X_0, \dots, X_M)}$$

$$P(p_A = p|X_0, \dots, X_M) = \frac{p^{m_A} (1-p)^{m_B} \cdot \mathbf{1}_{[0,1]}(p)}{Z}$$

Inférence Bayésienne - Exemple 1, tirage aléatoire

? On tire M boules, m_A de type A et m_B de type B quelle est la distribution $P(p_A = p | X_0, \dots, X_M)$?

$$P(p_A = p | X_0, \dots, X_M) = \frac{P(X_0, \dots, X_M | p_A = p) \cdot P(p_A = p)}{P(X_0, \dots, X_M)}$$

$$P(p_A = p | X_0, \dots, X_M) = \frac{p^{m_A} (1-p)^{m_B} \cdot \mathbf{1}_{[0,1]}(p)}{Z}$$

$$P(p_A = p | m_A, m_B) = \frac{\Gamma(m_A + m_B + 2)}{\Gamma(m_A + 1)\Gamma(m_B + 1)} p^{m_A} (1-p)^{m_B} \cdot \mathbf{1}_{[0,1]}(p)$$

$$\Rightarrow p_A | m_A, m_B \sim \beta(m_A + 1, m_B + 1)$$

Inférence Bayésienne - Exemple 2, moyenne d'une gaussienne

On observe N réalisations indépendantes de variables aléatoires $(X_0 = x_0, \dots, X_{N-1} = x_{N-1})$ de loi gaussienne, de variance connue σ^2 et de moyenne μ inconnue.

? Quelle est la distribution $P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1})$

Inférence Bayésienne - Exemple 2, moyenne d'une gaussienne

On observe N réalisations indépendantes de variables aléatoires $(X_0 = x_0, \dots, X_{N-1} = x_{N-1})$ de loi gaussienne, de variance connue σ^2 et de moyenne μ inconnue.

? Quelle est la distribution $P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1})$

$$P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) \propto \exp\left(-\frac{\left(m - \frac{\sum x_i}{N}\right)^2}{2\frac{\sigma^2}{N}}\right) P(\mu = m)$$

Inférence Bayésienne - Exemple 2, moyenne d'une gaussienne

On observe N réalisations indépendantes de variables aléatoires $(X_0 = x_0, \dots, X_{N-1} = x_{N-1})$ de loi gaussienne, de variance connue σ^2 et de moyenne μ inconnue.

? Quelle est la distribution $P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1})$

$$P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) \propto \exp\left(-\frac{\left(m - \frac{\sum x_i}{N}\right)^2}{2\frac{\sigma^2}{N}}\right) P(\mu = m)$$

⇒ Comment choisir la distribution *a priori* $P(\mu = m)$?

Inférence Bayésienne - Choix des distributions a priori

- ⇒ Comment choisir la distribution *a priori* $P(\mu = m)$?
- Distribution conjugée: loi a priori et loi a posteriori de la même forme
- ⇒ $P(\mu) = \mathcal{N}(\mu_0, \sigma_0)$, μ_0 est l'avis d'un expert et σ_0 le doute sur cet avis.

$$P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) \propto \mathcal{N}\left(\frac{\sum_i x_i \sigma_0^2 + \mu_0 \frac{\sigma^2}{N}}{\frac{\sigma^2}{N} + \sigma_0^2}, \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)$$

Inférence Bayésienne - Choix des distributions a priori

⇒ Comment choisir la distribution *a priori* $P(\mu = m)$?

- Distribution conjugée: loi a priori et loi a posteriori de la même forme

⇒ $P(\mu) = \mathcal{N}(\mu_0, \sigma_0)$, μ_0 est l'avis d'un expert et σ_0 le doute sur cet avis.

$$P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) \propto \mathcal{N}\left(\frac{\sum x_i}{N} \sigma_0^2 + \mu_0 \frac{\sigma^2}{N}}{\frac{\sigma^2}{N} + \sigma_0^2}, \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)$$

- Loi non informative dans la limite $\sigma_0 \rightarrow \infty$

⇒ $P(\mu) = 1$ (distribution impropre)

$$P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) = \mathcal{N}\left(\frac{\sum x_i}{N}, \frac{\sigma^2}{N}\right)$$

Inférence Bayésienne - Choix des distributions a priori

⇒ Comment choisir la distribution *a priori* $P(\mu = m)$?

- Distribution conjugée: loi a priori et loi a posteriori de la même forme

⇒ $P(\mu) = \mathcal{N}(\mu_0, \sigma_0)$, μ_0 est l'avis d'un expert et σ_0 le doute sur cet avis.

$$P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) \propto \mathcal{N}\left(\frac{\sum x_i}{N} \sigma_0^2 + \mu_0 \frac{\sigma^2}{N}}{\frac{\sigma^2}{N} + \sigma_0^2}, \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)$$

- Loi non informative dans la limite $\sigma_0 \rightarrow \infty$

⇒ $P(\mu) = 1$ (distribution impropre)

$$P(\mu = m | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) = \mathcal{N}\left(\frac{\sum x_i}{N}, \frac{\sigma^2}{N}\right)$$

Inférence Bayésienne - Exemple 3, variance d'une gaussienne

On observe N réalisations indépendantes de variables aléatoires

$(X_0 = x_0, \dots, X_{N-1} = x_{N-1})$ de loi gaussienne, de moyenne $\mu = 0$ et de variance $\sigma^2 = \frac{1}{\tau}$ inconnue.

Quelle est la distribution $P(\tau = t | X_0 = x_0, \dots, X_{N-1} = x_{N-1})$

Inférence Bayésienne - Exemple 3, variance d'une gaussienne

On observe N réalisations indépendantes de variables aléatoires

$(X_0 = x_0, \dots, X_{N-1} = x_{N-1})$ de loi gaussienne, de moyenne $\mu = 0$ et de variance $\sigma^2 = \frac{1}{\tau}$ inconnue.

Quelle est la distribution $P(\tau = t | X_0 = x_0, \dots, X_{N-1} = x_{N-1})$

$$P(\tau = t | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) \propto t^{N/2} \exp\left(-\frac{1}{2}t \sum x_i^2\right) P(\tau = t)$$

Inférence Bayésienne - Exemple 3, variance d'une gaussienne

On observe N réalisations indépendantes de variables aléatoires

$(X_0 = x_0, \dots, X_{N-1} = x_{N-1})$ de loi gaussienne, de moyenne $\mu = 0$ et de variance $\sigma^2 = \frac{1}{\tau}$ inconnue.

Quelle est la distribution $P(\tau = t | X_0 = x_0, \dots, X_{N-1} = x_{N-1})$

$$P(\tau = t | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) \propto t^{N/2} \exp\left(-\frac{1}{2}t \sum x_i^2\right) P(\tau = t)$$

Comment choisir la distribution *a priori* $P(\tau = t)$?

⇒ Comment choisir la distribution *a priori* $P(\tau = t)$?

- Distribution conjugée: loi a priori et loi a posteriori de la même forme

⇒ Il s'agit d'une loi gamma $\Gamma(k, \theta)(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$, k certitude expert en nombre de mesures équivalentes, $k\theta$ valeur estimée (on passe les calculs)

⇒ Comment choisir la distribution *a priori* $P(\tau = t)$?

- Distribution conjugée: loi a priori et loi a posteriori de la même forme
⇒ Il s'agit d'une loi gamma $\Gamma(k, \theta)(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$, k certitude expert en nombre de mesures équivalentes, $k\theta$ valeur estimée (on passe les calculs)
- Distribution non informative: on se place à la limite $k = 0, \theta \rightarrow \infty : P(\tau = t) = \frac{1}{t}$

Inférence Bayésienne - Exemple 3, variance d'une gaussienne

On observe N réalisations indépendantes de variables aléatoires

$(X_0 = x_0, \dots, X_{N-1} = x_{N-1})$ de loi gaussienne, de moyenne $\mu = 0$ et de variance $\sigma^2 = \frac{1}{\tau}$ inconnue.

? Quelle est la distribution $P(\tau = t | X_0 = x_0, \dots, X_{N-1} = x_{N-1})$

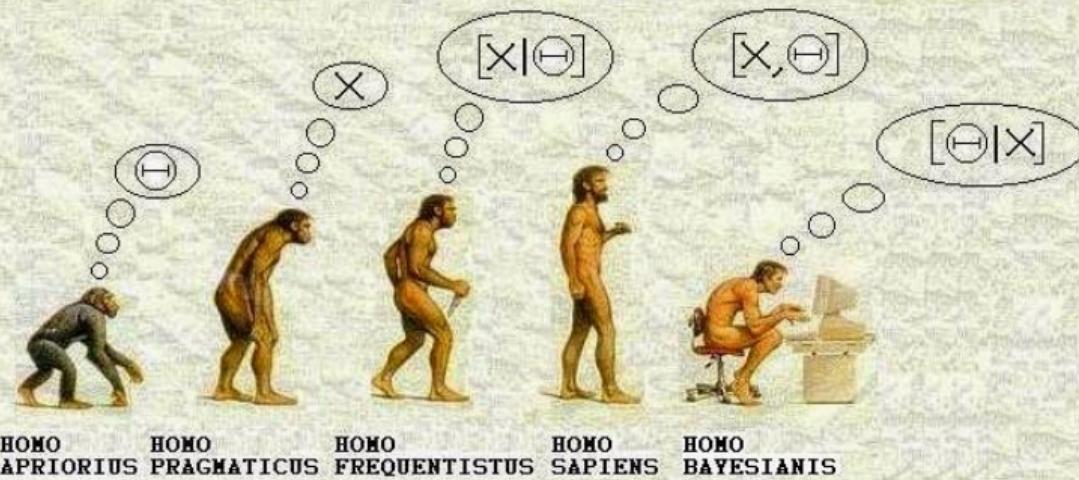
A priori impropre: $P(\tau = t) = 1/t$

$$P(\tau = t | X_0 = x_0, \dots, X_{N-1} = x_{N-1}) \propto t^{N/2-1} \exp\left(-\frac{1}{2}t \sum x_i^2\right)$$

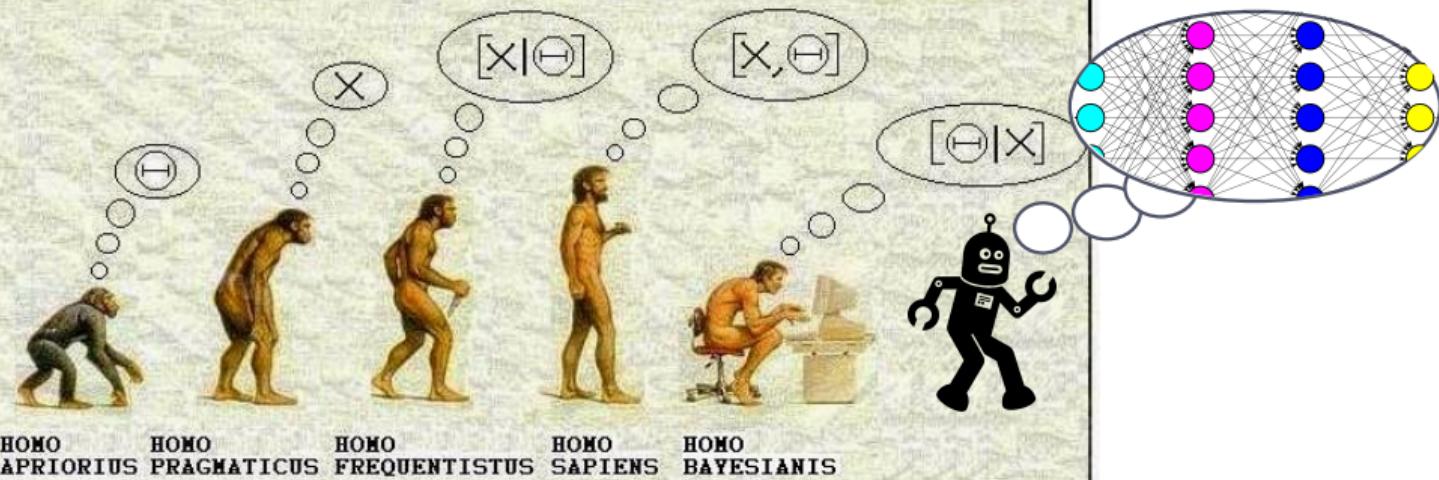
$$\Rightarrow \tau | (x_0, \dots, x_{N-1}) \sim \Gamma\left(\frac{N}{2}, \sum x_i^2\right)$$

$$\Rightarrow \sigma^2 | (x_0, \dots, x_{N-1}) \sim \text{Inv}\Gamma\left(\frac{N}{2}, \sum x_i^2\right) \text{ (Inverse Gamma)}$$

(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...



Apprentissage Machine

Algorithme paramétrique

Un *algorithme* est une procédure qui à partir de données x dans une espace \mathcal{E}_X produit des résultats y dans un autre espace \mathcal{E}_Y :

$$A : \mathcal{E}_X \rightarrow \mathcal{E}_Y$$

Algorithme paramétrique

Un *algorithme* est une procédure qui à partir de données x dans une espace \mathcal{E}_X produit des résultats y dans un autre espace \mathcal{E}_Y :

$$A : \mathcal{E}_X \rightarrow \mathcal{E}_Y$$

Notamment, deux types d'algorithmes:

Algorithme paramétrique

Un *algorithme* est une procédure qui à partir de données x dans une espace \mathcal{E}_X produit des résultats y dans un autre espace \mathcal{E}_Y :

$$A : \mathcal{E}_X \rightarrow \mathcal{E}_Y$$

Notamment, deux types d'algorithmes:

- Algorithmes de classification sur C classes \mathcal{E}_Y est un ensemble discret de taille C

Algorithme paramétrique

Un *algorithme* est une procédure qui à partir de données x dans une espace \mathcal{E}_X produit des résultats y dans un autre espace \mathcal{E}_Y :

$$A : \mathcal{E}_X \rightarrow \mathcal{E}_Y$$

Notamment, deux types d'algorithmes:

- Algorithmes de classification sur C classes \mathcal{E}_Y est un ensemble discret de taille C
 - Classification simple: \mathcal{E}_Y est l'ensemble des vecteurs de dimension C ,
 $y \in \mathcal{E}_Y \Rightarrow \forall i, y_i \geq 0, \sum_i y_i = 1$

Algorithme paramétrique

Un *algorithme* est une procédure qui à partir de données x dans une espace \mathcal{E}_X produit des résultats y dans un autre espace \mathcal{E}_Y :

$$A : \mathcal{E}_X \rightarrow \mathcal{E}_Y$$

Notamment, deux types d'algorithmes:

- Algorithmes de classification sur C classes \mathcal{E}_Y est un ensemble discret de taille C
 - Classification simple: \mathcal{E}_Y est l'ensemble des vecteurs de dimension C ,
 $y \in \mathcal{E}_Y \Rightarrow \forall i, y_i \geq 0, \sum_i y_i = 1$
 - Algorithmes de régression: \mathcal{E}_Y est un \mathbb{R} espace vectoriel

Algorithme paramétrique

Un *algorithme* est une procédure qui à partir de données x dans une espace \mathcal{E}_X produit des résultats y dans un autre espace \mathcal{E}_Y :

$$A : \mathcal{E}_X \rightarrow \mathcal{E}_Y$$

Notamment, deux types d'algorithmes:

- Algorithmes de classification sur C classes \mathcal{E}_Y est un ensemble discret de taille C
 - Classification couple: \mathcal{E}_Y est l'ensemble des vecteurs de dimension C ,
 $y \in \mathcal{E}_Y \Rightarrow \forall i, y_i \geq 0, \sum_i y_i = 1$
 - Algorithmes de régression: \mathcal{E}_Y est un \mathbb{R} espace vectoriel

Un *Algorithme paramétrique* est un algorithme A_θ dont la procédure dépend de paramètres regroupés dans un vecteur θ .

Apprentissage machine supervisé

L'apprentissage supervisé est un domaine qui permet:

Apprentissage machine supervisé

L'apprentissage supervisé est un domaine qui permet:

- de choisir une certaine classe d'algorithmes A_θ

Apprentissage machine supervisé

L'apprentissage supervisé est un domaine qui permet:

- de choisir une certaine classe d'algorithmes A_θ
- de régler les paramètres de θ selon une démarche rationnelle sur un ensemble de données *d'apprentissage annotées* $\{(x_i, y_i)\}$ dans une *phase d'apprentissage*

Apprentissage machine supervisé

L'apprentissage supervisé est un domaine qui permet:

- de choisir une certaine classe d'algorithmes A_θ
- de régler les paramètres de θ selon une démarche rationnelle sur un ensemble de données *d'apprentissage annotées* $\{(x_i, y_i)\}$ dans une *phase d'apprentissage*
- une fois θ réglé et fixé à θ_* , de *prédirer* les valeurs y pour de nouvelles données x , dans une *phase d'inférence*

Apprentissage machine supervisé

L'apprentissage supervisé est un domaine qui permet:

- de choisir une certaine classe d'algorithmes A_θ
- de régler les paramètres de θ selon une démarche rationnelle sur un ensemble de données d'apprentissage annotées $\{(x_i, y_i)\}$ dans une phase d'apprentissage
- une fois θ réglé et fixé à θ_* , de prédire les valeurs y pour de nouvelles données x , dans une phase d'inférence

La phase d'apprentissage se fait en général par maximisation d'une fonction d'adéquation entre les sorties de l'algorithme évalué sur les données d'apprentissage et les annotations:

$$\theta_* \in \operatorname{Argmin}_\theta \mathcal{L}((A_\theta(x_i)), (y_i)_i)$$

\mathcal{L} fonction de perte, minimale si adéquation maximale.

Régression logistique

Classifieur binaire probabiliste:

- Entrée = vecteur de dimension M ,

- Sortie = vecteur stochastique de dimension 2: $A_\theta(x) = \begin{pmatrix} A_\theta(x)_0 \\ A_\theta(x)_1 \end{pmatrix}$, $A_\theta(x)_i$, $i = 0, 1$

est l'estimation de $P(Y(X) = i|X)$.

Régression logistique

Classifieur binaire probabiliste:

- Entrée = vecteur de dimension M ,

- Sortie = vecteur stochastique de dimension 2: $A_\theta(x) = \begin{pmatrix} A_\theta(x)_0 \\ A_\theta(x)_1 \end{pmatrix}$, $A_\theta(x)_i$, $i = 0, 1$

est l'estimation de $P(Y(X) = i|X)$.

Hypothèse: $\forall i$, $\log P(X = x|Y(X) = i) = \sum_j \theta_j^{(i)} x_j + \text{Cste}$

ce qui conduit à $P(Y(X) = 0|X = x) = \frac{\exp \sum_j \delta_j x_j}{1 + \exp \sum_j \delta_j x_j}$, $\delta_j = (\theta_j^{(0)} - \theta_j^{(1)})$

Régression logistique

Classifieur binaire probabiliste:

- Entrée = vecteur de dimension M ,

- Sortie = vecteur stochastique de dimension 2: $A_\theta(x) = \begin{pmatrix} A_\theta(x)_0 \\ A_\theta(x)_1 \end{pmatrix}$, $A_\theta(x)_i$, $i = 0, 1$

est l'estimation de $P(Y(X) = i|X)$.

Hypothèse: $\forall i$, $\log P(X = x|Y(X) = i) = \sum_j \theta_j^{(i)} x_j + \text{Cste}$

ce qui conduit à $P(Y(X) = 0|X = x) = \frac{\exp \sum_j \delta_j x_j}{1 + \exp \sum_j \delta_j x_j}$, $\delta_j = (\theta_j^{(0)} - \theta_j^{(1)})$

Fonction de perte: vraisemblance des données $\mathcal{L}(A_\theta(x_i)), (y_i)_i) = \prod_i A_\theta(x_i)_{y_i}$

Régression logistique

Classifieur binaire probabiliste:

- Entrée = vecteur de dimension M ,

- Sortie = vecteur stochastique de dimension 2: $A_\theta(x) = \begin{pmatrix} A_\theta(x)_0 \\ A_\theta(x)_1 \end{pmatrix}$, $A_\theta(x)_i$, $i = 0, 1$

est l'estimation de $P(Y(X) = i|X)$.

Hypothèse: $\forall i$, $\log P(X = x|Y(X) = i) = \sum_j \theta_j^{(i)} x_j + \text{Cste}$

ce qui conduit à $P(Y(X) = 0|X = x) = \frac{\exp \sum_j \delta_j x_j}{1 + \exp \sum_j \delta_j x_j}$, $\delta_j = (\theta_j^{(0)} - \theta_j^{(1)})$

Fonction de perte: vraisemblance des données $\mathcal{L}(A_\theta(x_i)), (y_i)_i) = \prod_i A_\theta(x_i)_{y_i}$

Minimisation de \mathcal{L} par descente de gradient .

Descente de gradient

$f(\theta)$ à optimiser, on construit la suite (θ_i) avec

$$\theta_{i+1} = \theta_i - \mu \frac{\partial f}{\partial \theta}(\theta_i)$$

$\mu > 0$ est un paramètre à fixer.

⇒ Si μ bien choisi, (θ_i) converge vers un minimum local de f .

Descente de gradient

$f(\theta)$ à optimiser, on construit la suite (θ_i) avec

$$\theta_{i+1} = \theta_i - \mu \frac{\partial f}{\partial \theta}(\theta_i)$$

$\mu > 0$ est un paramètre à fixer.

⇒ Si μ bien choisi, (θ_i) converge vers un minimum local de f .

On peut exploiter les dérivées d'ordre 2:

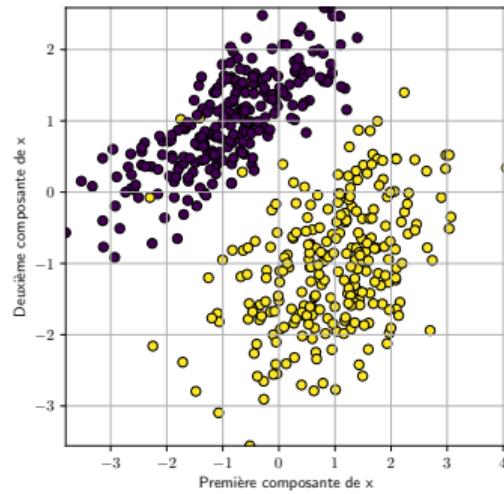
$$f(\theta + \delta\theta) \simeq f(\theta) + {}^t \frac{\partial f}{\partial \theta}(\theta_i) \cdot \delta\theta + \frac{1}{2} {}^t \delta\theta H \delta\theta$$

Si θ de dimension n H matrice de taille $n \times n$ (Hessienne), $H_{kl} = \frac{\partial^2 f}{\partial \theta_k \partial \theta_l}$

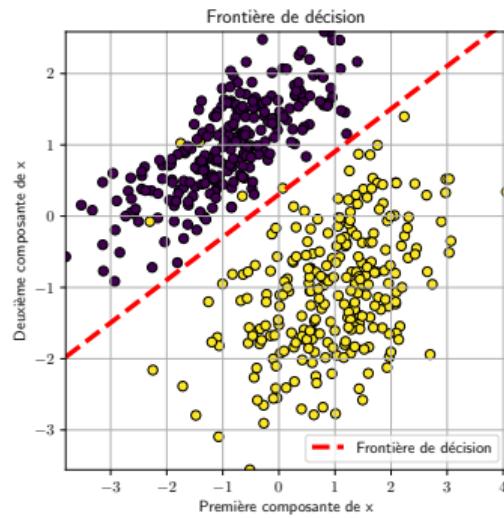
On prend

$$\theta_{i+1} = \theta_i - H^{-1} \frac{\partial f}{\partial \theta}(\theta_i)$$

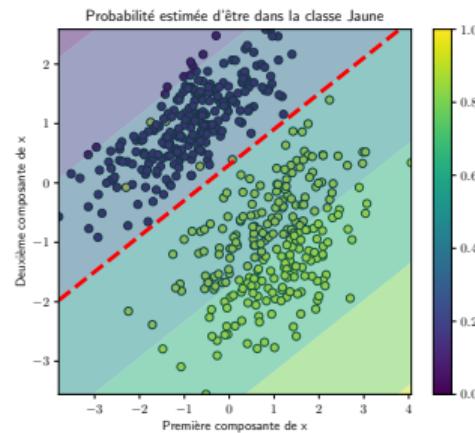
Régression logistique - Exemple



Régression logistique - Exemple

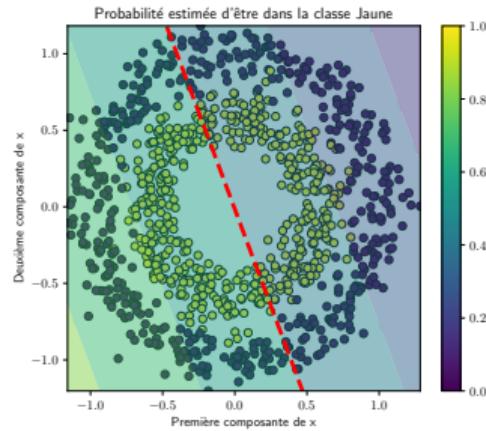
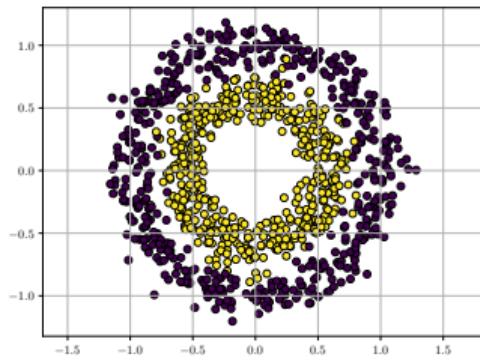


Régression logistique - Exemple



Régression logistique - Séparateur non linéaire

Si on applique une régression logistique sur des ensembles de points non séparables linéairement, le résultat est "décevant":

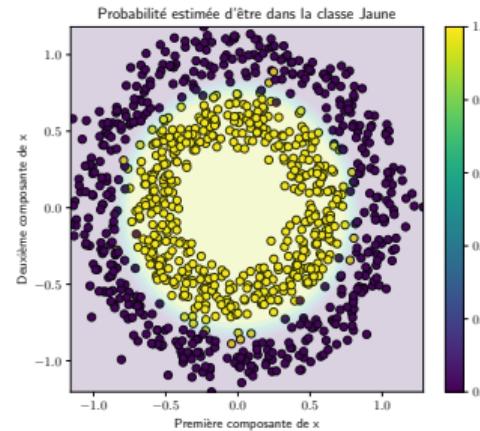


Régression logistique - Séparateur non linéaire

Une solution: plonger les données

- dans un espace de plus grande dimension
- avec une fonction non linéaire

Exemple ici: $t(x_0, x_1) \rightarrow t(x_0, x_1, x_0^2 + x_1^2)$



Le perceptron

Généralise la régression logistique.
Classifieur souple sur C classes.

Le perceptron

Généralise la régression logistique.

Classifieur souple sur C classes.

Hypothèse: $\log P(Y(X) = c|X = x) \propto S(\sum w_i^{(c)}x_i + b_i) = Q^{(c)}$

Avec S une fonction scalaire non linéaire, typiquement

$$S(x) = \text{sigmoïde}(x) = \frac{1}{1+\exp -x}.$$

Le perceptron

Généralise la régression logistique.

Classifieur souple sur C classes.

Hypothèse: $\log P(Y(X) = c | X = x) \propto S(\sum w_i^{(c)} x_i + b_i) = Q^{(c)}$

Avec S une fonction scalaire non linéaire, typiquement

$$S(x) = \text{sigmoïde}(x) = \frac{1}{1 + \exp -x}.$$

Au final, $P(Y(X) = c | X = x) = \text{softmax}((Q^{(k)})_k)(c)$

Le perceptron multicouches

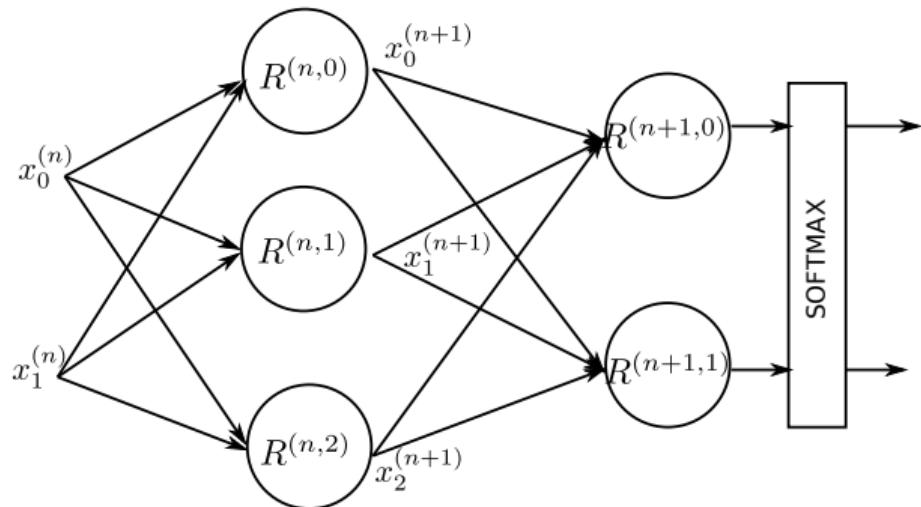
Classifieur souple sur C classes.

N couches, la n -ème = M_n perceptrons.

$x^{(n)}$ entrées de la n -ème couches,
 $x^{(0)}$ =les données d'entrée.

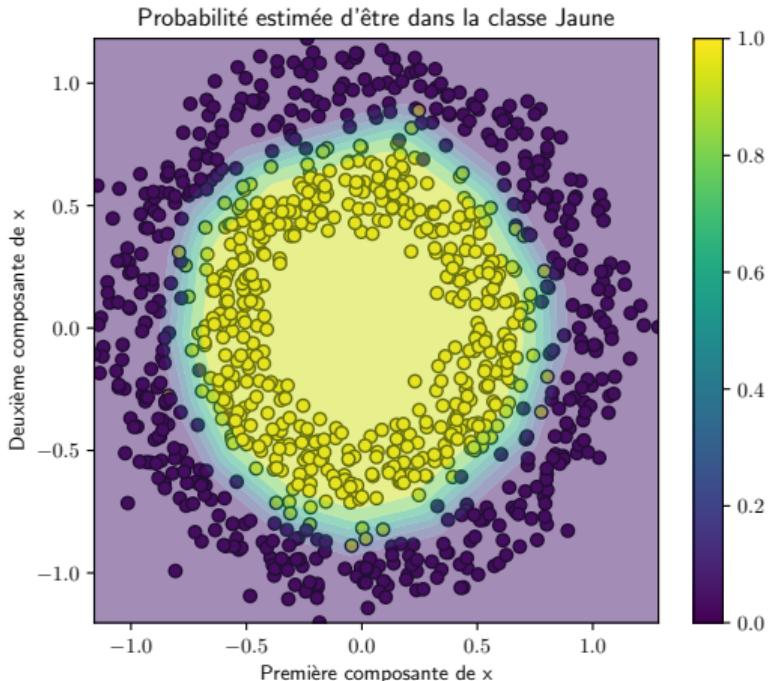
La n -ème couche renvoie M_n sorties $R^{(n,m)}(x^{(n)}) = S(\sum w_k^{(n,m)} x_k^{(n)} + b_k^{(n,m)})$

Finish: fonction softmax

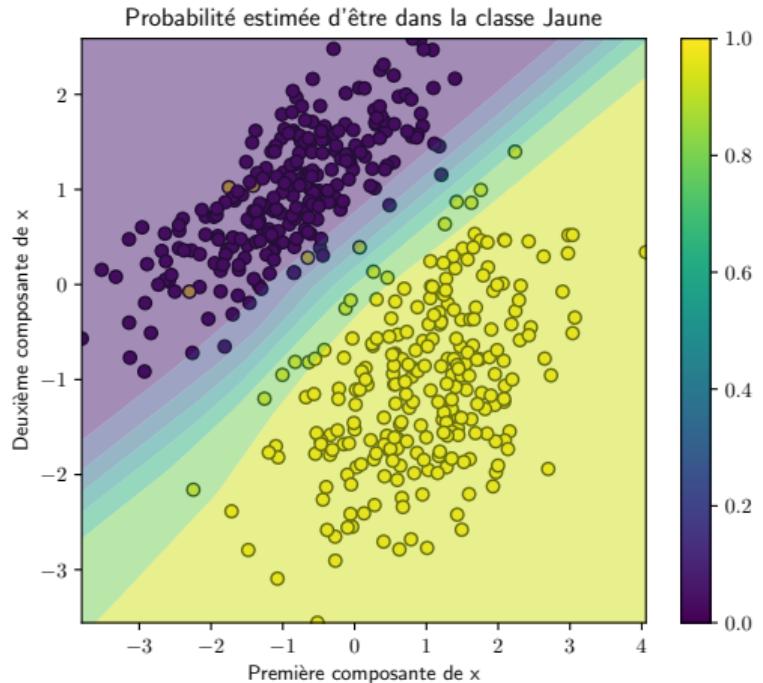


Optimisation par une implémentation particulière de la descente du gradient:
l'algorithme de rétro-propagation (voir cours deep learning)

Perceptron multicouches - résultats



Perceptron multicouches - résultats



Algorithmes de classification avec apprentissage supervisé

Il y a beaucoup d'autres algorithmes !

- Support Vector Machines
- Arbres
- Forêts aléatoires
- AdaBoost
- Gradient Boosting

Algorithmes de classification avec apprentissage supervisé

Il y a beaucoup d'autres algorithmes !

- Support Vector Machines
- Arbres
- Forêts aléatoires
- AdaBoost
- Gradient Boosting

La bibliothèque Python Sklearn (Scikit-Learn) regroupe l'implémentation de bon nombre d'entre eux.

Prototype SKlearn

Scikit-learn est une bibliothèque Python dédié à l'apprentissage machine, qui implémente notamment de nombreux algorithmes d'apprentissage supervisé

```
class MLClassifier:  
    def fit(X,y): # Phase d'apprentissage  
        # X shape (n_samples, n_features)  
        # y shape (n_samples) of integers  
        # return _  
        ...  
    def predict(X):  
        # X shape (n_samples, n_features)  
        # return y shape (n_samples) of integers  
        ...  
        return y  
    def predict_proba(X):  
        # X shape (n_samples, n_features)  
        # return y shape (n_samples,n_classes) of floats  
        ...  
        return y  
    def predict_log_proba(X):  
        # X shape (n_samples, n_features)  
        # return y shape (n_samples,n_classes) of floats  
        ...  
        return y
```